University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Doctoral Dissertations 1896 - February 2014

1-1-2004

# Using performance level descriptors to ensure consistency and comparability in standard setting.

Dafter January Khembo
*University of Massachusetts Amherst*

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation
Khembo, Dafter January, "Using performance level descriptors to ensure consistency and comparability in standard setting." (2004). *Doctoral Dissertations 1896 - February 2014*. 2375.
https://scholarworks.umass.edu/dissertations_1/2375

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST
312066 0289 0072 5

# USING PERFORMANCE LEVEL DESCRIPTORS TO ENSURE CONSISTENCY AND COMPARABILITY IN STANDARD SETTING

A Dissertation Presented

by

DAFTER JANUARY KHEMBO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2004

Department of Education, Policy, Research, and Administration

© Copyright by Dafter J. Khembo 2004

All Rights Reserved

USING PERFORMANCE LEVEL DESCRIPTORS TO ENSURE CONSISTENCY AND COMPARABILITY IN STANDARD SETTING

A Dissertation Presented

by

DAFTER JANUARY KHEMBO

Approved as to style and content by:

Ronald K. Hambleton, Chair

Stephen G. Sireci, Member

Lynn Shelley-Sireci, Member

Andrew Effrat, Dean
School of Education

To my patient and loving wife, Rosemary

ACKNOWLEDGEMENTS

My grateful thanks are due to the Malawi Government for organizing funding and giving me the rare opportunity to study for the Doctor of Education degree at the University of Massachusetts.

I would also like to thank members of my committee, which comprised Professors Ronald Hambleton (Chair), Stephen G Sireci and Lynn Shelley-Sireci, for their prompt helpful comments and advice each time I turned to them for guidance.

The completion of this dissertation would not have been possible without the encouragement and assistance I obtained from the management and fellow members of staff of the Malawi National Examinations Board. Special thanks are due to Mr M. W. Matemba, the Executive Director of MANEB, who ensured that I got everything I needed to complete my study. Mention is made of the financial assistance the Board made available to me to supplement my research budget.

I would also like to thank all the subject matter experts who accepted to participate in this study. They traveled long distances and stayed away from their homes for five days for the sake of this project.

Thanks are also due to my brothers in-law Alexander and Moses Dossi for their moral and financial assistance rendered to my family while I was away doing my doctoral studies.

Finally I would like to thank my wife, Rosemary, and children: Ausbert, Loness, Felix, and Mtisunge, for their support and patience during my long absence from them.

Thank you all.

ABSTRACT

USING PERFORMANCE LEVEL DESCRIPTORS TO ENSURE CONSISTENCY AND COMPARABILITY IN STANDARD SETTING

MAY 2004

DAFTER JANUARY KHEMBO, B.Ed, UNIVERSITY OF MALAWI

M.A. (EDUCATION), UNIVERSITY OF LONDON INSTITUTE OF EDUCATION

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

The need for fair and comparable performance standards in high-stakes examinations cannot be overstated. For examination results to be comparable over time, uniform performance standards need to be applied to different cohorts of students taking different forms of the examination.

The motivation to conduct a study on maintenance of the Malawi School Certificate of Education (MSCE) performance standards arose following the observation by the Presidential Commission of Enquiry into the MSCE Results that the examination was producing fluctuating results whose cause could not be identified and explained, except for blaming the standard setting procedure that was in use. This study was conducted with the following objectives: (1) to see if use of performance level descriptors could ensure consistency in examination standards; (2) to assess the role of training of judges in standard setting; and (3) to examine the impact of judges' participation in

scoring students' written answers prior to being involved in setting examination standards.

To maintain examination standards over years means assessing different cohorts of students taking different forms of the examination using common criteria. In this study, common criteria, in the form of performance level descriptors, were developed and applied to the 2002 and 2003 MSCE Mathematics examination, using the item score string estimation (ISSE) standard setting method. Twenty MSCE mathematics experts were purposely identified and trained to use the method.

Results from the study demonstrated that performance level descriptors, especially when used in concert with test equating, can help greatly determine grading standards that can be maintained from year to year by reducing variability in performance standards due to ambiguity about what it means to achieve each grade category. It has also been shown in this study that preparing judges to set performance standards is an important factor for producing quality standard setting results. At the same time, the results did not support a recommendation for judges to gain experience as scorers prior to participating in standard setting activities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

All examinations used for certification must set a passing score that distinguishes certifiable from not certifiable examinees. Examinees achieving the passing score are judged to possess the minimal knowledge and skills necessary for the award of a certificate. However, in certificate examinations, such as the Malawi School Certificate Examination (MSCE) and Massachusetts Comprehensive Assessment System (MCAS), the examinees may be classified into more than pass and fail categories. In such cases, instead of using a single passing score, psychometricians use multiple *cut scores* or *grade boundaries*, which sort examinees into categories that reflect different levels of proficiency. Other names for cut score[1] are: standard, achievement level, threshold level, minimum proficiency level, mastery level (Hambleton, 2001) or cut score. The process of deriving the cut scores is called *standard setting* (Cizek, 1996). Thus, the cut score, which represents the minimum proficiency for a performance category, is the numeric outcome of a standard setting process. Examinees who are assigned to a particular category are assumed to have met the minimum requirements for that level (Kane, 2001).

One of the reasons for using examinations for making certification and admission decisions is because they are believed to be fair to all examinees. Fairness is required not only within the same group of examinees but across cohorts as well. In other words, this year's certified candidates should not be held to a higher or lower standard than the standard applied to previous years' candidates (Johnson, Squires, & Whitney, 2002),

[1] The terms cut score, cutoff score, standard, passing score, minimum proficiency level, threshold level, mastery level, and grade boundary are used interchangeably in this study.

unless standards have been purposely changed. In order to achieve fairness across cohorts, the same standards need to be applied over time and for different forms of the examination. This is one of the major challenges facing examining institutions: to ensure that standards remain the same over time. As certificates awarded to candidates every year are valued the same, it is important to ensure that similar grades on the certificate represent the same level of proficiency. This can be assured by applying consistent standards when grading different cohorts of examinees. Thus, a cut score for the same grade could be numerically different due to differences in test difficulty, but reflecting the same level of ability. By applying consistent standards every year, the certificates awarded would be comparable in terms of what the grades on the certificates indicate what the recipients know and can do. If the standards applied to grade the examinees are not comparable, then the meaning of the certificate is unclear (Norcini, 1997).

One way to ensure fairness and grade comparability is to develop examinations of equivalent difficulty and maintain the same cut scores from one year to the next. However, although test developers try to construct assessments of equivalent difficulty, it turns out to be hard to do (Angoff, 1971; Newton, 1997) because of a finite number of items available to build a test, and sometimes an absence of valid and stable field-test item statistics.

The most common way practitioners ensure comparability of standards is by test score equating. This means, for example, that some items are common to successive forms of the examination. (Other equating designs are possible, but common item non-equivalent groups design is the most popular.) The performance on the common items is used to estimate the relative level of ability in the groups taking the examination. The

actual cut score is adjusted as a result of this in order to maintain standards over forms of a test. Because of security difficulties, the Malawi National Examinations Board (MANEB) does not reuse items. In fact, many testing organizations prefer not to reuse test items. This makes statistical equating difficult to carry out in practice.

If examinees and the curriculum are more or less consistent, and because it is difficult to build parallel forms of the examination, then another way to maintain standards would be to set cut scores on subsequent examinations to produce consistent results. For example, if 50% of examinees passed the examination last year, then set cut score on this year's examination to allow 50% to pass. The problem with this approach is that it does not allow for some growth or drop in achievement to be reflected in the results if changes in the quality of examinees take place. Therefore, fixing pass rate is not acceptable. Thus, if it is not possible to maintain standards over forms and time by test score equating and fixing pass rate, then other methods have to be tried.

This study addresses the question of whether the use of performance level descriptors can help ensure consistency, comparability, and fairness in the determination of cut scores for the MSCE Mathematics test. Descriptions of the knowledge and skills necessary to achieve particular grade categories were developed. Once such knowledge and skills are clearly explained to a standard setting panel, the judges can use these descriptions to set cut scores on different forms of the examination. Since the same performance descriptions will be used for each version of the examination, these cut scores will reflect the same achievement levels. Or, judges can use these descriptors to equate the scores of subsequent examination papers to the achievement scale, thus ensuring consistent standards are employed from year to year (Bennett, 1998). It would

then be possible to compare performance of different groups who have taken different forms of the examination. Of course, it remains to be demonstrated that such an approach would actually work.

### 1.1.1 Characteristics of the Examination Investigated

The MSCE Mathematics examination consists of two subtests, known as Paper 1 and Paper 2 (see Appendices A to D for 2002 and 2003 MSCE Mathematics examination papers). By design, Paper 1 is constructed to be easier than Paper 2, although the papers are weighted the same: each paper carries 100 marks (score points). Examinees take both. Paper I consists of 24 compulsory questions, and time allowed to answer the questions is two hours. Paper 2, which has two sections, is allocated $2\frac{1}{2}$ hours. Section A consists of six compulsory questions and is worth 55 marks. Section B has also six questions, but candidates choose any three to answer. Each question in section B is worth 15 marks. All the questions in Paper 2 have two parts.

### 1.1.2 Processing of Examination Results

The processing of examination results begins with the scoring of examinees' answers. After the examinations have been administered the answer booklets for all examinees are taken to a central place called a *marking center* for scoring. The scorers, called Assistant Examiners, are trained in scoring before they are allowed to score. The Chief Examiners or Senior Assistant Examiners supervise the scoring activities.

Prior to the main scoring, the Chief Examiners pre-score some answers to familiarize themselves with the responses. This affords them the opportunity to consider

some of the students' correct responses, which may not have been included on the provisional model answers prepared by the examination developers.

The main scoring begins with the standardization of the model answers. During this time, more alternative solutions are discussed and the marking scheme is usually expanded to allow for a greater variety of answers than had been anticipated. The purpose of standardization, therefore, is to ensure that all possible correct responses are properly recognized, and that all scorers score to the same standard.

While scoring is in progress, all already scored answer booklets are checked by Script Checkers who look for errors that the Assistant Examiners might have made, such as wrong additions of the subtotals, responses that may not have been scored, and wrong transfer of scores.

The Chief Examiner also checks up to 10% of each Assistant Examiner's work. Erratic scoring can be detected at this stage, and appropriate measures to correct the situation are taken. Examinees' total scores for the paper are entered on a score sheet, which the Data Entry Clerks use for entering the scores into the computer. Another Data Entry Clerk verifies the entered scores. In general, it would appear that the quality control measures are sufficiently stringent to keep the margin of error arising from scorers' mistakes within bounds.

During standard setting, known as *awards meeting* in Malawi, summary statistics in the form of mean, standard deviation, mode, minimum score, maximum score, and frequency distribution are used to aid the award process, which is discussed in the next section.

### 1.1.3 Grade Awarding Process

In Malawi, students taking the MSCE examination have a choice of 22 subjects (See Appendix E for a full list of the subjects). Each subject is graded on a nine-point scale:

1-2, denote pass with distinction;

3-6, denote pass with credit;

7-8, denote general pass; and

9 denotes fail

To be awarded an MSCE certificate, candidates must pass any combination of at least six subjects, including English, with at least one credit grade. The certificate can also be awarded if a candidate passes five subjects, including English, with three credit grades.

The grade awarding process entails converting examinees' scores into grades. The awards committee uses cut scores, commonly known as grade boundaries, to turn scores into grades. The important people at the awards meeting are: The Executive Director of MANEB who chairs the award panel; the Chairperson, who is a subject specialist who gives background information about the examination papers in relation to test development; the Chief Examiner, who is also a subject matter specialist who supervises the scoring in his/her subject, and provides most of the information regarding the validity of the examination papers and their comparability to previous papers; the Subject Officer, who is a MANEB officer who services the awards meeting in his/her subject; and the Research and Test Development Officer, who is another MANEB officer who computes

intermediate cut scores in each performance category, and provides the panel with statistical information in the form of impact data.

When converting examinees' scores into grades, the Chief Examiner suggests to the panel, without reference to examination statistics, what he/she thinks the cut scores for 2/3 (distinction/credit), 6/7 (credit/pass), and 8/9 (pass/fail) should be. To do this, he/she uses his/her past experience as well as his/her experience while scoring the examinee answers. Where a subject has more than one paper, each paper is graded separately. Then the corresponding cutoff scores at 2/3, 6/7, and 8/9 in all the papers are summed to arrive at the final cut scores for the subject.

### 1.1.4 Grading Assumptions

The grading process makes the assumptions that: the examinations are equivalent across years in terms of difficulty level, content covered, and skills examined; the test administration conditions are uniform from year to year; and the cohorts taking the examination each year are randomly equivalent.

These assumptions imply that approximately the same pass rates should be expected from one year to the next, since logically, students from one year to the next are assumed to be just about equivalent. It follows, therefore, that similar grade distributions would be expected every year. This argument is in line with the definition given by Cresswell (1996):

> Two successive examinations on the same syllabus have comparable standards if two groups of candidates who attend the same schools receive grades which are identically distributed after studying the syllabus and taking the examinations. (p. 83)

Kane (2001) concurs with Cresswell and argues further that in the absence of any events that would cause a sharp change in the competence of the candidates the results of a new standard-setting study would be expected to yield similar results. If a new cut score produces a significantly different pass rate, the appropriateness of one or both of the cut scores would be suspect.

## 1.2 Statement of the Problem

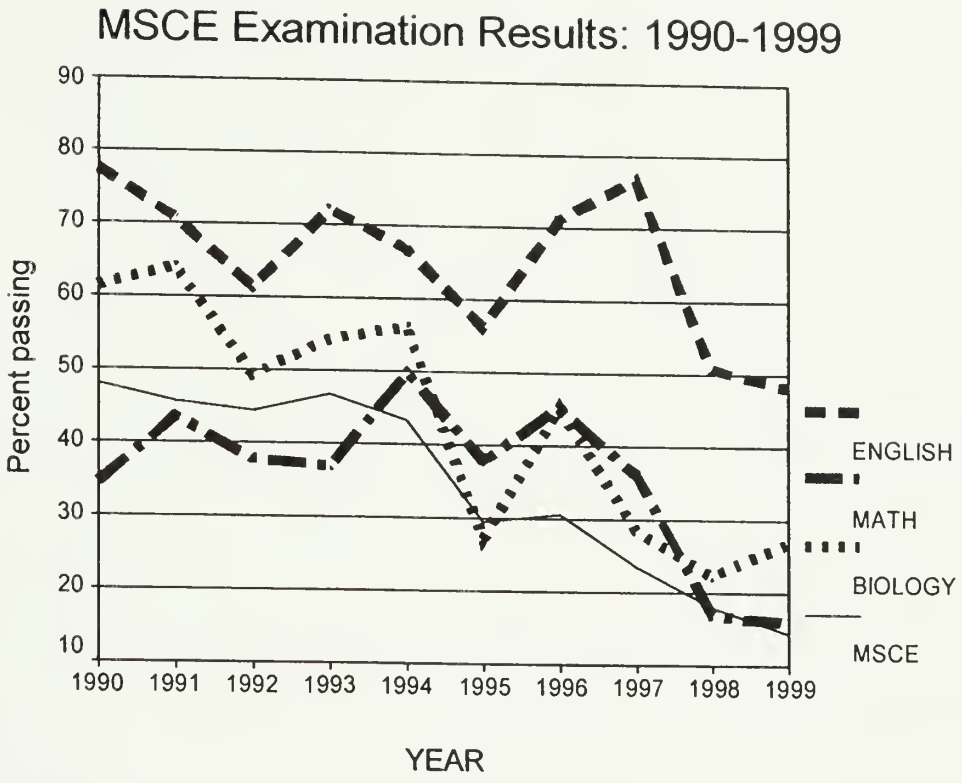
The logical consequence of the grading assumptions cited above would be equivalent pass rates from year to year. However, fluctuating pass rates and a generally downward trend in the percentage of examinees passing the examination have been observed as shown in Table 1.1 and Figure 1.1.

Table 1.1 MSCE Results: 1990-1999

| Year | English | Math | Biology | MSCE |
|---|---|---|---|---|
| 1990 | 77.5 | 34.3 | 61.4 | 48.0 |
| 1991 | 70.7 | 43.5 | 64.0 | 45.5 |
| 1992 | 61.3 | 37.7 | 49.3 | 44.4 |
| 1993 | 72.0 | 36.9 | 54.3 | 46.7 |
| 1994 | 66.8 | 49.8 | 56.0 | 43.1 |
| 1995 | 56.2 | 37.9 | 27.0 | 29.4 |
| 1996 | 71.1 | 45.2 | 45.2 | 30.7 |
| 1997 | 76.8 | 36.1 | 28.4 | 23.6 |
| 1998 | 50.7 | 16.9 | 22.3 | 17.9 |
| 1999 | 48.1 | 15.8 | 26.8 | 14.3 |

Source: Malunga et al., 2000.

Figure 1.1




Based on these results, there was a big drop in 1992 and 1995 in all three subjects. There was a sharp rise in 1996. In 1997 only English pass rate improved. There was another sharp drop in 1998 in all the subjects. While minor variations would be expected, some of the large fluctuations shown in Table 1.1 and Figure 1.1 are not easy to explain. The 1999 results caused a public outcry, which prompted the government to institute a Commission of Enquiry to investigate the causes of the poor results. The MSCE Commission of Enquiry apportioned some blame on the grading process:

> It is unlikely that a change of such magnitude could be explained either by changes in the examination population or by a deterioration in the quality of teaching and learning...The large changes in the apparent performance of candidates, both from one year to the next in individual subjects and from one subject to another in any given year, must have been considered by MANEB. In particular, when grade thresholds are decided for a given subject and the standards for that subject are set, those responsible must have been aware that the outcome would be a large change in the proportion of candidates passing the examination compared to the previous session. (pp. 45-46)

This is precisely the reason that led to the decision to conduct this study.

As already stated, the grading assumptions outlined above would mean equivalent pass rates from year to year. The difficult question is: Why are the MSCE results not consistent with the grading assumptions? The fluctuating pass rates mean that at least one of the variables is not constant as assumed. If they are constant, then the fluctuating results can only be explained by changing standards, otherwise only minor fluctuations would be observed. If standards vary over time as it has been shown above, then there is some degree of unfairness (Mathews, 1985), because there is lack of equity, consistency and uniformity. So, what is the way out?

### 1.2.1 Purpose of the Study and Research Questions

When examinations are used for certification as the MSCE is, the important assumption underlying the use of a cut score is that it is an accurate discriminator of mastery and non-mastery in the content domain (Heubert & Hauser, 1999). However, there is always an estimation error attached to each cut score (Tanner, 1996). This error may be due to the particular selection of judges, the vargaries of applying a method, the qualifications of the judges, etc.

The goal of this study was to develop a standard setting method that would increase people's confidence in the comparability of grades awarded in different years. To do this, the study developed explicit criteria in the form of performance level descriptors, representing what the examinees should know and be able to do in order to be classified in a particular grade category. The logic of performance level descriptors is that they remain the same from year to year, regardless of differences in test difficulty. If these performance level descriptors are applied every year as criteria for grading students, then the annual variability due to the ambiguity about what it means to achieve each level of proficiency will be reduced, and the cut scores for different forms of the examination will represent about the same attainment level. It was believed that this will increase people's confidence in grade consistency and comparability. In addition, this will result in fairness to all cohorts of examinees since they will be graded using the same standards.

The study was designed, therefore, to answer the following questions:

1. What knowledge and skills should students demonstrate in order to be graded fail, pass, credit, or distinction?
2. How would the standard setting results of two sub-panels using the same performance level descriptors compare?
3. Does the application of the same performance level descriptors yield consistent results over years?
4. How do the equated cut scores that are based on common items compare with those that are based on common judges?
5. How do the SMEs' ratings before and after scoring students' answers compare?

6. How do the standards set by trained SMEs compare with those set by untrained SMEs using the same performance level descriptors?

### 1.2.2 Justification of the Study

The declining and fluctuating pass rates for MSCE examinations have been blamed on a number of factors, including the standards set by MANEB (Malunga et al., 2000). It must be admitted that MANEB test developers do try to develop examinations that are approximately equivalent in difficulty. Guidelines are available for this purpose. However, as it has already been noted, it is extremely difficult to develop tests that are exactly equivalent in difficulty (Angoff, 1971; Newton, 1997), even with the aid of statistical procedures. This means that different cohorts of students would take tests that are somewhat different in difficulty level. This becomes an issue of fairness. For an assessment system to meet the test of fairness it must be consistent in its demands on the students. It is therefore necessary to ensure equity: different cohorts should be examined using uniform standards.

It should be noted that the present standards were established a long time ago, when the examination was first administered in 1972, and a lot of things have changed since then. For example, the syllabi for the various MSCE subjects have been changing as need for change in the subjects arose. The major change occurred in 1996 following a secondary school curriculum review. It is recommended that standards be reset whenever there have been major curricular revisions or item format changes (Hambleton, 2000). MANEB has also made a number of changes in the assessment methods. For example, the school-based assessment has been removed from a number of subjects such as

Agriculture and Geography. Further, for logistical reasons, the numbers of examination papers have been reduced in most subjects. For example, mathematics used to be assessed through three papers, namely Algebra, Arithmetic, and Geometry. Now it is assessed through two papers. Item format has also changed in papers of some subjects. Changes have also occurred in terms of administration conditions. For example, unlike in the past, the police are now involved in the administration of the examination. In addition, from 2003, candidates have been taking the examination in cluster centers, not in their schools, as has been the practice before. Furthermore, the fact that MSCE certificate is now the minimum requirement for employment in the civil service (JCE used to be the minimum requirement) means that student motivation towards the examination is not the same. It means that there is greater pressure to pass the examination now than before. Because of the changes that have taken place as outlined above, it is appropriate for Malawi to consider alternative standard-setting methods that are suitable for its present assessment system.

## 1.3 Context of the Problem

To better understand and appreciate the concerns this study intends to address, some contextual information is helpful. To this effect, this section describes the geographical location of Malawi, its history, education and examination systems.

### 1.3.1 Where is Malawi?

Malawi is located to the eastern side of Southern Africa. It lies between latitudes 9 and 17 degrees south of the equator, and between longitudes 33 and 36 degrees east. It

shares border with Mozambique to the south and east, Zambia to the west, and Tanzania to the east and north.

Malawi is 901 kilometers long and ranges in width from 80 to 161 kilometers. It is 118,486 square kilometers in size, of which 94,276 square kilometers are land area. The remaining area is composed of Lake Malawi, which is about 475 kilometers long. The Great African Rift Valley runs the entire length of Malawi and passes through Lake Malawi down to Shire Valley. The Shire River drains the water from Lake Malawi into the Zambezi River in Mozambique.

About 75% of land is plateau area lying between 1400 m and 3077 m above sea level. The lower area lies as low as 1000 m above sea level. These varied features account for varied climatic conditions. Rainfall and temperatures vary depending on altitude and proximity to the lake. The southern end of Malawi, which is low lying and close to the sea, has high temperatures. The highlands have cooler temperatures.

There are generally three seasons: a cool dry season from May to August; a hot dry season from September to November; and a rainy season between November to April. Annual rainfall ranges between 800 mm to 2000 mm. In general, the climate is relatively favorable for diversified Agriculture which accounts for 85% of Malawi's exports, with tobacco, tea, and sugar being the major export commodities.

The country is divided into three administrative regions. The Northern Region is sparsely populated and comprises 11% of the country's population. The Central Region is moderately populated with about 40% of the population. The Capital City, Lilongwe, is located in the Central Region. The Southern Region is more densely populated with nearly 50% of the country's population. Most of the country's commercial activities take

place in this region. The commercial capital, Blantyre, and the Old Administrative Capital, Zomba, where MANEB is, are situated in this region.

### 1.3.2 A Brief History of Malawi

The Malawi nation comprises people who migrated from other regions within and outside Africa. These groups of people formed the Maravi Kingdom which operated until 1891 when Britain established Nyasaland Protectorate. On 23rd October, 1953, Britain combined Nyasaland with the Federation of Northern and Southern Rhodesia (now Zambia and Zimbabwe respectively), and the three countries were ruled under the Federation of Rhodesia and Nyasaland.

Following pressure to end the Federation, a new constitution was agreed upon in 1960, and Malawi held first elections under a new constitution in 1961. The federation was formally dissolved in 1963 (Tindall, 1992), and in the same year, the territory was granted self-government. In 1964, Malawi gained independence, and attained a Republican status in 1966. The new constitution established a one-party state, and opposition movements were suppressed. Malawi was therefore under one-party rule until 1993 when Malawian voters in a referendum and rejected the one-party state. In 1994 Malawians voted in a multiparty general elections, and a new government was formed. At the time of writing this dissertation Malawians were preparing for the third multiparty general elections.

### 1.3.3 Malawi's Education and Examination Systems

Malawi's education system, which borrows heavily from the British system, consists of a pyramidal structure of 8 years of primary education, 4 years of secondary education, and 4 years of tertiary education with a broad primary base, narrowing down to relatively small enrolments at the secondary and tertiary levels. Public examinations serve as certification and selection devices at the end of primary and secondary cycles. Primary School Leaving Certificate Examination (PSLCE) and MSCE examination are administered to graduates of primary and secondary schools respectively. Another public examination, Junior Certificate Examination (JCE), is required for continuation after Form 2 at secondary school. These are all certification examinations. In addition, PSLCE and MSCE are also used for admission to secondary school and university or other post-secondary institutions respectively. Furthermore, MSCE has now become the minimum qualification for employment in civil service. Until recently JCE used to be the minimum requirement for civil service employment. MANEB is responsible for developing and administering all these examinations.

### 1.3.4 History of MSCE Examination

Before Malawi attained independence from Britain in 1964, school leavers (secondary school graduates) were taking the Overseas School Leaving examination offered by the University of Cambridge Local Examination Syndicate (UCLES) of the United Kingdom (UK). In 1968 it was decided to localize the examination. Consequently, in 1969, the Malawi parliament enacted a law that created the Malawi Certificate Examination (MCE) Board. This Board was charged with the responsibility of

developing and administering the Malawi Certificate of Education (MCE) examination with the assistance of the Associated Examining Board (AEB) of the UK. MCE examination was first introduced in June 1972. At that point in time, all examination papers were still set and scored in England. Gradually the responsibility of setting and scoring the examination was transferred from the British to the Malawian Chief and Assistant Examiners. Later, the MCE Board changed its name to MCE & Testing Board (MCE & TB), because it began providing other testing services such as aptitude testing, besides examining students. The MCE & TB continued to administer MCE examinations with the AEB until 1989 when the handover was completed. The MCE examination became known as the Malawi School Certificate of Education (MSCE) examination, and is considered to be equivalent to the O-level examination of the UK.

Following an evaluation of examinations in Malawi in 1984, it was recommended that all school examinations be developed and administered by one central authority (William, 1984). Consequently, in 1987, parliament approved legislation merging the examinations section of the Ministry of Education with the MCE & TB, thus forming the Malawi National Examinations Board (MANEB), which currently operates all school examinations. The Ministry of Education used to administer JCE and PSLCE examinations before then.

# CHAPTER 2

# REVIEW OF THE LITERATURE

## 2.1 Introduction

This literature review is organized into three major sections. The first section gives general information on standard setting. The section begins by defining standard setting, followed by a discussion of the importance of standard setting in examinations. A history of standard setting concludes the section. The second section describes various standard setting methods in both the test-centered and examinee-centered categories. The process of developing performance level descriptors that guide a standard setting process in performance assessments is also covered. The concept of misclassification errors (false positives and false negatives) is addressed as are guidelines for running a standard setting study and how it can be evaluated. The third and final section of the chapter presents findings from some standard setting studies. The studies are in four categories: (1) those that investigated the importance of training participants in a standard setting study; (2) those that are about maintenance of examination standards over time; (3) those that looked at standards set by different panels; and (4) those that investigated standards of the MSCE examinations. Since there is a vast literature on this topic, this review is not exhaustive, and readers are referred to other relevant literature in the main text.

## 2.2 General Information on Standard Setting

### 2.2.1 What Is Standard Setting?

The word *standard* has multiple meanings. According to the *Webster's Universal College Dictionary* (2001), the word, in one sense, means *normal* or *usual.* In another

sense, it means something that others of a similar type are compared to or measured by. In yet another related sense, it means a rule or principle that is used as a basis for judgment. It is the last two meanings that are of interest in this study. They imply the level of quality or excellence by which the actual qualities of individuals are judged. In the context of educational assessment, a standard is an explicit decision rule that assigns each examinee to one of several categories of performance based on his or her test score (Cohen, Kane, & Crooks 1999, cited by Reckase, 2001). This means that examinees of one category are more similar to one another than are examinees of different categories.

The process of arriving at a *passing score,* which classifies examinees into pass or fail categories is called s*tandard setting* (Cizek, 1996). Because in certificate examinations, examinees can be classified into more than pass and fail categories, Cizek (2001) revised the definition to the "task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons will be made" (p. 3). The points on a score scale that demarcate the levels of performance are variously called standards, achievement levels, passing scores, minimum proficiency levels, threshold levels, mastery levels, cut scores or cutoff scores (Hambleton, 2001). These cut scores serve the purpose of sorting examinees into two or more categories that reflect different levels of proficiency.

At this point, a distinction needs to be made between *performance standards* and *content standards* as used in education. Content standards are curricular frameworks (also known as objectives) that specify what should be taught at each grade level, while performance standards refer to the various levels of proficiency that the examinees are

expected to demonstrate in relation to the content standards (Linn, 1995; Hambleton, 2001).

Another distinction should also be made between *conceptual standards* and *operational standards*. A conceptual standard is the concept in a person's mind that enables that person to decide whether something is or is not good enough, while an operational standard is a rule for deciding whether something is or is not good enough.

> A person's conceptual standard refers to all those proficiencies that the person considers relevant for classifying the student... An operational standard refers to only those proficiencies that are actually measured, usually by some kind of test. Therefore, operational standards often take a form of cut scores. (Livingston, 1995, p. 39)

### 2.2.2 The Importance of Standard Setting in Examinations

According to Zieky (2001), "there are many situations ... in which cut scores are mandated by law and people have no choice but to set them" (p. 25). For example, examining institutions have the legal authority to give tests, set standards, and change them (Kane, 2001).

One of the reasons for setting standards is to help in making informed decisions. In education, for example, there are decisions to be made about rewarding merit, allocating resources in ways that maximize cost-benefit (Cizek, 2001), and screening examinees (Zeiky & Livingston, 1977). Relevant information is necessary to help make wise decisions. Information derived from standard setting can be used for making certification, accountability, and categorization decisions (Linn, 1995). Other uses of standard setting information include:

1. Giving meaning to scores from a test or examination: For example, in norm-referencing, "This student scored in the top 15% of students"; Or in criterion-referencing, "This student scored above the mastery score on this test" (Cresswell, 2001), or "More students scored in the *credit* category this year than last year"
2. As blueprints: by specifying what students are expected to learn, content standards provide blueprints for what is important to teach as well as to test (Messick, 1995). Hansche (1998) concurs with Messick and says: "For students, the expectations outlined in content and performance standards provide a framework for understanding what they need to know and be able to do to meet the requirements for each performance level... Students who understand what is expected are more likely to feel ownership of their own progress toward meeting the standards... For teachers, content standards provide a broad framework to help them focus on the curriculum and what is most important for students to learn." (p. 11)

### 2.2.3 History of Standard Setting

One of the purposes of some testing programs is to identify those who should be declared to have passed or failed. Standard setting is required to determine the point on the score scale beyond which the candidates are deemed to have passed and below which they are deemed to have failed. Passing and failing can be likened to acceptance and rejection of an individual for succeeding or failing to perform an activity, respectively. Instances of acceptance and rejection are available in the Holy Bible[2], implying that

[2] Biblical quotations are from *The Holy Bible,* New International Version, Michigan: Zondervan Bible Publishers, 1986. Zieky quoted from *The Holy Scriptures*, New York: Hebrew Publishing Co., 1939.

standard setting is as old as the Bible. Zieky (1995) quotes the following two biblical stories as examples of standard setting:

> Jephthah then called together the men of Gilead and fought against Ephraim. The Gileadites struck them down because the Epraimites had said, "You Gileadites are renegades from Ephraim and Manasseh." The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, "Let me cross over," the men of Gilead asked him, "Are you an Ephraimite?" If he replied, "No," they said, "All right, say 'Shibboleth.' " If he said, "Sibboleth," because he could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan. Forty-two thousand Epraimites were killed at that time. (Judges 12: 4-6)

This story exemplifies how performance standards work. According to the Gileadites, ability to pronounce the word "Shibboleth" was the performance standard that Giliedites are able to perform, but not the Ephraimites. The ability to pronounce the word distinguished the Gileadites from the Ephraimites, just like a cut score distinguishes masters from non-masters.

Another biblical story that relates to standard setting is to be found in Genesis 18:22-23. In the story, Abraham had learned that God was planning to destroy the city of Sodom. Concerned, Abraham asked God:

> Will you sweep away the righteous with the wicked? What if there are fifty righteous people in the city? Will you really sweep it away and not spare the place for the sake of fifty righteous people in it? Far be it from you to do such a thing – to kill the righteous with the wicked, treating the righteous and the wicked alike…The lord said, "If I find fifty righteous people in the city of Sodom, I will spare the whole place for their sake. Then Abraham spoke up again: "…What if the number of the righteous is five less than fifty? Will you destroy the whole city because of five people?" "If I find forty-five there, "He said, :I will not destroy it." God and Abraham continued to discuss in the same manner until God agreed to spare Sodom for the sake of ten righteous people. (Genesis 18: 22-32)

This story also demonstrates the process of setting cut scores. In this case, the cut score was set at 10, after some discussion (due process requirement) between God and Abraham. The story also demonstrates how difficult it is to completely avoid misclassification errors – false positives and false negatives - in standard setting. In this story, some righteous people would be misclassified as unrighteous, if there were only less than 10 of them in the city. Similarly, in educational testing situations, some people who should pass can fail and those who should fail can pass. This problem is described in detail later under "Misclassification Errors"

Away from the Bible, some of the earliest recorded testing programs were conducted in China. Around 2000 and 200 B.C., China conducted military selection and civil service examinations, respectively (Cizek, 2001). This means that people started performing standard setting exercises as early as 2000 B. C.

Before Nedelsky introduced modern ways of determining cut scores about half a century ago, cut scores were determined based on tradition e.g. 'The passing score is 70% because it has always been 70%.' In some cases, the determination of a cut score was a one-person decision: people in power would decide what a passing score should be (Zieky, 1987). In selection examinations, the number of places available determined where the cut score would be set. For example, in England, a cut score for secondary school entrance examination was determined by drawing a line at the requisite point, decided by the number of secondary school places actually available, and declaring those below the line to have failed (Sutherland, 1984, quoted by Zieky, 1995). This practice is still operational in the modern world as acknowledged in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999):

> But when tests are used for selection, it may be appropriate to rank-order examinees according to their test performance and establish a cut score so as to select a prespecified number ... (p. 50).

## 2.3 Standard Setting Methods

Since Nedelsky (1954) proposed the first systematic standard setting method, several other psychometricians have come up with alternative methodologies for determining cut scores for examinations. Most of the newly developed methods were intended to overcome the weaknesses in the earlier methods.

In just over 30 years after the development of Nedelsky's (1954) method, Berk (1986) documented 38 new standard setting methods. Other notable standard setting methods include: *minimally acceptable person* (Angoff, 1971); *minimally qualified applicant* (Ebel, 1979); *judgmental policy capturing (JPC,* Jaeger, 1995); *borderline group* (Zeiky & Livingston, 1977); *contrasting-groups* (Zeiky & Livingston, 1977); *direct consensus* (Pitoniak, Hambleton, & Sireci, 2002); *item score string estimation* (Impara & Plake, 1998); *cluster analysis* (Sireci, 2001); *body of work* (Kingston, Kahl, Sweeney, & Bay, 2001); and *bookmark method* (Mitzel, Lewis, Patz, & Green, 2001). These methods have been classified into test-centered and examinee-centered categories. These are described next.

### 2.3.1 Test-Centered Methods

Test-centered methods require participants[3] to review the items or tasks on the test and decide on the level of performance on these items or tasks required to meet the

[3] The terms participant, judge, rater, subject matter expert (SME), and panelist are used interchangeably in this study.

performance standard. Some examples of test-centered standard setting methods are given below.

#### 2.3.1.1 Nedelsky Method

Nedelsky's (1954) F-D student method requires subject matter experts or instructors to read each item, and identify the alternatives that a minimally competent candidate (called the F-D student by Nedelsky, where F denotes failure and D denotes barely passing) should reject as being incorrect. The rationale for this approach is the assumption that minimally competent candidates eliminate wrong answers that they can identify as wrong, and then guess among the remaining alternatives. The reciprocal of the remaining options is the probability that an F-D examinee would get the item correct. The sum of these probabilities of all the items on the test is the examinee's expected score. This expected score of borderline candidates is the basis for the cut score.

The major shortfall of Nedelsky method is that it can only be used with multiple-choice items. In addition, Berk (1984) quoted by Cizek, (1996) observed that the Nedelsky scale does not permit values between 0.5 and 1.0. Shepard (1980) noted that the judges who were using the Nedelsky method were reluctant to assign a probability of 1.0, which would mean a 100% chance of getting an item correct. According to Shepard, it was for this reason that the Nedelsky method produced standards which were lower than those produced by other methods.

### 2.3.1.2 Angoff Method and its Derivatives

The Angoff (1971) method is like the Nedelsky method in that, in its basic form, it is used in dichotomously scored items, and requires a judge to make a judgment about the probability that a minimally acceptable person will get the item correct. Unlike the Nedelsky method, the Angoff method does not require the judges to eliminate the obviously incorrect choices on each item. The method requires the judges

> ... to state the probability that a 'minimally acceptable person' would answer each item correctly.... The sum of these probabilities, or proportions, would then represent the minimally acceptable score (Angoff, 1971, p.515).

One of the advantages of the Angoff method is its flexibility, which has resulted in the development of the "modified Angoff method." Several modifications to the approach have been developed to accommodate polytomously scored items. For this purpose, Hambleton and Plake, (1995) developed the "extended Angoff" method, in which the judges estimate the score that a borderline examinee would earn on each question. Then, estimates from all the judges on a question are averaged. These averages are then summed to get the cut score. If necessary, the per-question averages can be weighted to reflect their importance, difficulty, or some other attribute. The method is simple, and allows the judges to differentially weight the questions according to the importance they attach to the individual questions. However, concerns have been raised about the atomistic nature of the method, which might ignore the holistic nature of the performance being assessed.

Loomis and Bourque (2001) described four other Angoff derivative methodologies.

The percent correct method. The method requires judges to estimate the percentage of students who would write a response that would at least earn a partial credit. The major weakness of the method is that, while it is able to distinguish an incorrect response from partially or completely correct responses, it fails to distinguish between partially and completely correct responses. This raises reliability concerns.

Proportional method. In this method, judges estimate the percentage of students at the borderline who could write a response scored at each response point. The method takes into account partial credit.

Mean estimation method. In this method, judges estimate the average score for each polytomous item for students performing at the borderline.

In spite of its flexibility advantage, some concerns have been raised against the Angoff method. In their report: *Setting performance standards for student achievement,* Shepard and associates (1993) concluded that the Angoff method was "fundamentally flawed" (p. 151), because judges were inconsistent in the application of the method: they recommended lower cut scores when using easy items, and higher cut scores when using difficult items. Green (2000) also noted that "making probability estimates is a demanding task, and people don't do it very consistently" (p. 3). These alleged shortcomings came after Berk (1986) analyzed the technical adequacy of Nedelsky, Angoff, and Ebel standard setting methods, from which he concluded that the Angoff method was more technically adequate and easy to use than the other two. In agreement with Berk, Mehrens (1995), after reviewing a number of studies, reported that:

> The review of the literature suggests the general acceptance of the Angoff method as the preferred model, and this is my recommendation. The recommendation is based on the general reasonableness of the standard set, the ease of use, and the psychometric properties of the standard. (p. 231)

Kane (1995) also defended the Angoff method and issued rebuttals to three of the studies on which Shepard, et al. (1993) based their criticisms of the Angoff method:

> It is not clear that the other three studies point to any serious limitations in the Angoff procedures. The evidence developed in the five studies of the technical properties of the 1992 NAEP standard setting do not seem to justify the conclusion (Shepard, et al., 1993) based largely on these studies... (p. 129).

Also in defense of the Angoff method, Hambleton et al. (1999) issued rebuttals to the Pellegrino et al. (1999) report which had attacked the method used by NAEP. Hambleton and his colleagues observed that the conclusions of the Pellegrino et al. study were influenced by prejudice, and the study did not adhere to scientific principles.

#### 2.3.1.3 Item Score String Estimation Method

This method requires judges to give a *yes* or *no* answer if the borderline students can answer the dichotomously scored item correctly or incorrectly, respectively. The National Assessment for Educational Progress (NAEP) pilot tested the method during the 1994 achievement-levels setting (Loomis and Bourque, 2001). However, NAEP advisors expressed reservations about the method and recommended that the method be abandoned. Impara and Plake (1998) used the method and reported that "the experts are able to (a) conceptualize the minimally competent examinee group by identifying the skills and achievement levels that define this group and (b) make performance estimates

for this examinee group" (p. 70-71). The method can also be used for polytomously scored items by requiring judges to estimate the score borderline students would earn on each item. For a 3-point item, for example, judges would estimate whether a borderline examinee would score 1, 2, or 3.

#### 2.3.1.4 Ebel's Method

Ebel's (1979) method is one of the five standard setting procedures that he proposed. In fact, what is known as Ebel's method was not designed to be a method in its own right, but as a way of overcoming a weakness in another method.

> The second weakness of this approach can be overcome to some degree by determining the passing percentage from a subjective analysis of the relevance and difficulty of each item in the test. (p. 339)

According to the method, judges are required to read each item and make two classification decisions about it. First, the judges must classify the item by its relevance (essential, important, acceptable, or questionable). Then the judges must classify the item by its difficulty level (easy, medium, or hard). This produces a 3 x 4 matrix. The judges locate each item in its proper cell, according to their judgments. The judges then estimate the percentage of items in each cell of the matrix that a minimally qualified applicant should be able to answer correctly. The passing score is the sum of the products of the expected percentage correct in each category and the number of questions in that category.

One important advantage of the method is that it can be used with any item format. However, some concerns have been raised about the method. For example, Cizek

(1996) wondered how a test can contain a questionable item in the first place. Cizek also wondered why judges should be asked to estimate item difficulty when item analysis data can provide that information.

#### 2.3.1.5 Judgmental Policy Capturing (JPC) Method

All the methods described above assume that the tests to which they will be applied are unidimensional, and consequently, the items that compose the tests contribute to a summative scale. A newly developed method, the two-stage judgmental policy-capturing (JPC, Jaeger, 1995) utilizes several dimensions of performance. The standard setting participants are provided with a framework for developing acceptable and unacceptable profiles of performance across the dimensions. The judges study the examinees' score profiles, each consisting of a set of exercises or tasks. They classify each exercise using a pre-determined score-scale. To determine the judges' standard setting policies, a mathematical model (e.g. a linear regression model) is fit to their ratings to analyze their classifications. The analysis produces the distribution of importance weights that each judges attributes to each of the exercises. After discussing the results, the judges proceed to the second phase of standard setting, which involves rating the whole profile of each examinee, using another pre-determined score-scale that encompasses all dimensions of the attribute being assessed. The analysis of the ratings produces the overall performance of the candidates.

The major advantage of the method is that it allows very dissimilar information to be used to make decisions about examinees, and permits these to be weighted differently. However, the technique appears complicated to apply in practice.

#### 2.3.1.6 Direct Consensus Method

A newly developed method called direct consensus (Pitoniak et al., 2002) involves judges working with the actual exam scale. Judges set a cut score directly based on a description of master examinees, content of the examination, its scoring rubrics, statistical data that may be available, the judgments of other judges about the cut score, and a sample of candidates' constructed responses. Items are organized into sets or clusters based on content considerations. Cut scores are set on each cluster and then summed to obtain a cut score on the full exam. The facilitator engages the judges in a discussion of all the available information and attempts to help judges reach a consensus on the cut score. The goal in this method is to have the panel arrive at a cut score directly that they can agree with. In case of disagreements, the mean of their recommended cut scores is considered. It is generally a faster method than others available.

#### 2.3.1.7 Bookmark Method

As described by Mitzel et al. (2001), the bookmark method involves arranging test items in order of their difficulty, beginning with the least difficult, as determined by item response theory (IRT) calibrations. The judges review each item, and are asked to determine the knowledge, skills, and abilities that should be applied in order to correctly answer the item. The judges further determine what makes each item progressively more difficult than the previous one. Polytomously scored items appear multiple times in the ordered booklet, once for each score point, and the judges discuss the skills and knowledge required for each score point. The authors of the method recommend at least

three rounds of discussion, with the first round concentrating on identifying skills a given item requires for mastery. Judges may consider content that the examinees should master during rounds 2 and 3.

Judges are then asked to conceptualize an examinee at the threshold of a performance level (a barely proficient student). Keeping this examinee in mind, they mark a point in the booklet that they think represent the amount of material an examinee would need to master. Each judge presents his/her results, and discussion follows. The discussion forms the basis for second round of book-marking. In case of disagreements, the median cut score among the judges is calculated and adopted.

During round 3, judges consider impact data based on provisional round 2 results. Adjustments are suggested and considered in relation to the impact data, requisite skills, and content. If consensus is not reached, medians are calculated. Round 3 concludes with the final cut scores and impact estimates.

One major advantage of the bookmark method is that it is based on actual students' results in the sense that the ordering of items is based on students' performance, and that impact data is used when deciding on the final cut score. It has another advantage: it can handle both dichotomously and polytomously scored items. The use of IRT-calibrated items implicitly means that the method can account for the inter-item correlations (Kiplinger, 1997). Furthermore, the judges do not decide the difficulty of the items. The items are ordered in terms of difficulty level based on actual students' performance. Since content and skills are considered in the process of determining cut scores, the procedure can be a tool for evaluating what the test actually measures. The procedure gives information on whether there is anything missing from the test, or

whether there is something that should not be there. This, in turn, can lead to the construction of a more valid test (Kiplinger, 1997). One of several problems to the procedure is that it requires use of an appropriate IRT model, without which the accuracy of results becomes questionable.

#### 2.3.1.8 Cluster Analysis Method

The rationale of this method is based on the understanding that standard setting is a classification problem, where examinees are classified into categories based on their test scores such that test takers within each category are similar to one another, but different from test takers in other categories. According to the developer of the method, Sireci (2001), cluster analysis is one of the techniques for evaluating classification exercises, of which standard setting is one. As such, the cluster analysis method "is probably most useful for supplementing data derived from other standard-setting studies… or for evaluating standards already set on a test" (p. 340). If subject matter experts are available to interpret the clustering results, cluster analysis can also be used as a stand-alone standard setting method.

In any case, judgment is required to identify useful clusters, such as borderline or contrasting groups. In the case of a borderline cluster, the median score could be used as a passing score. For the contrasting groups, one could be an acceptable cluster and the other unacceptable clusters.

The method has a number of advantages which include applicability to polytomously scored items, ability to handle multidimensional data, non-reliance on panelists, ability to discover borderline and contrasting groups, etc. However, Sireci

admits that his method suffers from some weaknesses such as the need for test data before it can be applied, and that all examinees must respond to the same items, in which case, it is difficult to apply the method in examinations which offer examinees choice of questions to attempt.

### 2.3.2 Examinee-Centered Methods

In examinee-centered methods, performances of real examinees are evaluated relative to the performance standard (Jaeger, 1989). This section describes some notable methods from this category.

#### 2.3.2.1 Borderline Method

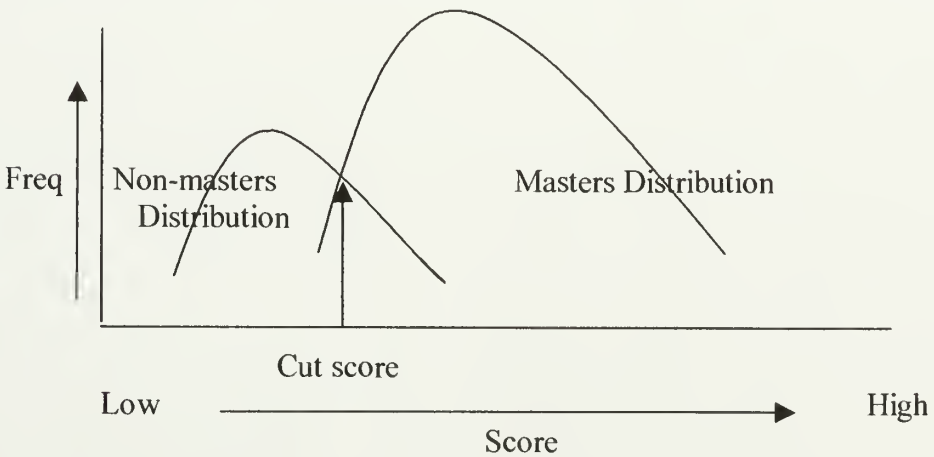
Zieky and Livingston (1977) proposed this procedure which requires judges first to identify the group of examinees whom they would classify as borderline (separating masters from non-masters) on the knowledge and skills assessed by the test. The placement of examinees in the borderline group is based on auxiliary information that is not related to the test. The median score of the borderline group can be chosen as the estimate of the standard, and it becomes the cut score. The major weakness of this method is that it is often difficult to find a sufficiently large group of borderline examinees (Cizek, 1996). Zieky and Livingston recommend at least 100 borderline examinees.

2.3.2.2 Contrasting Groups Method

Zieky and Livingston (1977) also proposed the *contrasting groups method.* In this method, like in the borderline, the focus is on the competences of the examinees rather the difficulty of the test. A group of teachers familiar with the examinees, and with the definitions of the groups into which the examinees are to be placed, separate the examinees into masters and non-masters groups based on their observations of the examinees in their classrooms. After the examinees have taken the examination and their answers have been scored, the score distribution for each group can be plotted on the same continuum. Where the distributions of the two groups meet (See Figure 2.1) is where the cut score between the two groups is set, since it is at this point that the classification errors are minimal. One concern about this method is the fallibility of judgments used to assign examinees to groups. Nevertheless, it is a relatively easy technique to implement and is easily understood by educators and parents.

Figure 2.1

An Illustration of Cut Score Determination Using the Contrasting Groups Method



#### 2.3.2.3 Body of Work Method

This method requires judges to examine complete student response sets, referred to as rich body of student work, and match them to performance level categories based on previously agreed on descriptions of what examinee at the different levels should know and be able to do (Kahl, Crockett, DePascale, & Rindfleisch (1993), cited by Kingston, Kahl, Sweeney & Bay, 2001).

The advantage of this method is that all of the information about an examinee is used to set standards, which is a more logical decision for standard setters to make. Discussions are focused on tangible examinees rather than intangible percentages of examinees passing test items.

### 2.3.3. Other Categorization Dimensions

Berk (1986) documented 38 different methods of standard setting methods. Many new methods have been developed since 1986, especially for polytomously-scored items. These methods have been classified in a variety of ways. This section describes some of them.

#### 2.3.3.1 Normative (Relative) and Absolute

According to Johnson, et al. (2002), normative (or relative methods as they are sometimes known) base the cut score on individual's rankings within a group. When the scores are high the cut score will be higher than when scores are low (Beuk, 1984). Relative standard setting methods as described by Nedelsky (1954), define adequate achievement by a student relative to his/her class or to any other particular group of

students. For example, a passing score can be set at one standard deviation below the mean for the group. It has the advantage of maintaining the pass rate over time. However it does not take into account changes in the students' achievement levels over time. For example, some able students may not pass the examination because there are so many high quality candidates in that particular pool. Conversely, if the pool of candidates is not particularly able, some students who have not attained a minimal level of proficiency may pass. Therefore, normative methods are not suitable for making decisions about proficiency.

In absolute methods, a student's performance is judged based on what constitutes an adequate level of achievement. The cut score is not dependent on the actual achievements of the participating examinees, but on an external criterion. Decisions about performance are linked to some pre-determined criteria for acceptable achievement.

#### 2.3.3.2 A Priori and a Posteriori

Another classification of standard setting methods reflects on the time of determining the cut scores. When cut scores are determined before test administration, the methods are described as *a priori*. These methods are generally based on judgments about the difficulty of the test items for a certain group of individuals. The Angoff method is a good example of this sub-category. When cut scores are determined after test data have been collected the standard setting methods are known as *a posteriori* (Gonzalez & Beaton, 1994). The cluster analysis method is an example.

#### 2.3.3.3. Constructed-Response and Selected-Response

Standard setting methods have also been classified on the basis of whether they are dealing with tests composed of selected-response items, constructed-response items, or both. The original forms of Angoff, Nedelsky, and Ebel methods are examples of selected-response methods, while extended Angoff, and other newly developed performance standard setting methods could be classified as constructed-response methods.

#### 2.3.3.4 Unidimensional and Multidimensional

Unlike the other methods, which assume unidimensionality of tests, the Jaeger's (1989) JPC method utilizes several dimensions of performance. The characteristics assessed are assumed to be complex performances, and the judges possess holistic notions of acceptable performance. The judges study the candidates' score profiles and rate them on a number of dimensions of the attribute being assessed. The cluster analysis standard setting method would be suitable for multidimensional data.

#### 2.3.3.5 Holistic and Analytic Methods

Holistic methods assume that achievement is highly integrated, and its "essential elements cannot be broken down into a series of small, independent tasks without destroying the essential meaning of performance" (Kane, 1995, p. 121). Examples of this category are body of work (BoW) method (Kahl, Crockett, DePascale, & Rindfleisch (1993), cited by Kingston, et al. 2001) and JPC (Jaeger, 1989). Analytic methods, on the other hand, assume that achievement can be assessed using relatively small parts of the

overall performance as indicators of achievement. The final cut score is derived by summing ratings from the individual tasks. The test-centered methods fall in this category. But the recent shift of emphasis to complex, performance-based assessments seem to point in the direction of holistic methods of standard setting.

#### 2.3.3.6 Compensatory and Conjunctive Methods

In compensatory methods, the overall judgment of the quality of performance depends on the average quality of all attributes (Coombs, 1964, quoted by Jaeger, 1995). This means that a high score on one scored dimension of an exercise can compensate for a low score on another. Methods that involve summing item cut scores are essentially compensatory methods. In conjunctive methods, the overall judgment depends on the quality of the weakest attribute. The examinees have to meet the minimum performance requirements in all attributes.

#### 2.3.3.7 Compromise Methods

Compromise methods (Beuk, 1984; Hofstee, 1983) were developed to strike a balance between relative and absolute methods. The Beuk's method involves asking the standard setting judges to make judgments about the minimum level of knowledge required to pass an examination, expressed as a percentage of the total raw score on the test. The judges are further asked to make another judgment about the passing rate expected, expressed as a percentage of the examinee population. After the examination has been administered, these expectations can be compared with reality. If there are differences, a compromise can be reached by making adjustments.

Hofstee's method recognizes political and cognitive considerations when deciding cut scores. Consequently, the implementation of Hofstee's method requires the judges to answer four questions related to these considerations: (i) What is the lowest cut score ($k_{min}$) that would be acceptable, even if every examinee passes? (ii) What is the lowest acceptable cut score, even if no examinee passes ($k_{max}$)? (iii) What is the maximum tolerable failure rate ($f_{max}$)? (iv) What is the minimum acceptable failure rate ($f_{min}$)? To derive a cut score, the points ($k_{min}$, $k_{max}$) and ($f_{max}$, $f_{min}$) are used to plot a line which is projected onto the distribution of observed scores. The intersection point will show the passing rate.

## 2.4 Performance Standard Setting

Performance assessments differ from other paper-and-pencil tests in that they require examinees to construct responses to a wide range of problems. Performance assessment is variously labeled alternative assessment – to distinguish it from traditional multiple-choice testing - or authentic assessment – to highlight the real world nature of tasks and contexts that make up the assessment. Whatever term is used, performance assessment implies active student production of evidence of learning, unlike multiple-choice, which is essentially passive selection among pre-constructed answers.

Several of the standard setting methods described above are not applicable to performance assessments. Some of them can be modified (e.g. Angoff method) to accommodate performance assessments. In some cases, new standard setting methods have been developed for performance assessments. Hambleton, et al. (2000) describes several of them.

According to Hambleton et al. (2000), performance standards are important in a number of ways including (a) providing a frame of reference for understanding test results; (b) providing more interpretive information about the meaning of test scores by defining performance categories; and (c) promoting excellence in education. Also commenting on the role of performance standards in reporting results in a more meaningful way, Haertel (2002) said:

> It is all but meaningless to say that a score of X or higher means "proficient" if all that can be said of "proficient" examinees is that they scored X or higher. Conversely, it may be quite valuable to have a substantive description of what it is that a given score indicates an examinee knows or can do. (p. 17)

### 2.4.1 Developing Performance Level Descriptors

The task of setting standards on performance assessments will basically involve two major steps: developing performance level descriptors and estimating cut scores (Kane 2001). Performance level descriptors are statements about what students need to know and be able to do to meet the requirements for a particular performance level. This sub-section describes how these performance level descriptors are developed.

The process of developing performance level descriptors begins with the determination of the number of performance levels and creating names for them. In most cases, this would be the responsibility of the examining agency. In other cases, this task is given to the standard setting judges. Yet, in other cases, the task is accomplished through a public hearing or consensus by educators (Thorn et al, 1990). When the performance levels have been named (or labeled), there remains a task of defining the meaning of

performance levels (Kingston et al., 2001). These policy definitions of performance levels are general and apply across all content domains (Hansche, 1998).

Building on these definitions, subject matter experts (SMEs) develop performance level descriptors for their subject areas. They develop descriptions for each performance level by considering and analyzing the general definitions of the levels. These descriptions are linked directly to content standards, which are assumed to already exist. (Content standards are statements of what students should know and be able to do.) Table 2.1, which was developed using information from Allen et al. (1997), describes how these terms should be understood as used by the National Assessment of Educational Progress (NAEP). The descriptors may be revised if they produce unacceptable results.

Hansche (1998) described the process of developing performance level descriptors as follows: The process starts with the determination of the foundation for developing performance standards, that is, how the standards will be used. The drafting of performance levels and performance descriptors follows this. The next step is to administer the assessments that are based on content standards and draft performance descriptors. It must also be determined how it will be known that students are meeting the standards. The performance descriptors might be revised if they produce unreasonable results. According to her, the process is iterative, and is based on the assumption that strong content standards already exist.

Table 2.1 NAEP Performance Levels Descriptions for Grade 4 Mathematics

| Performance level | Definition | Descriptor |
|---|---|---|
| Basic | This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade | Fourth-grade students performing at the basic level should show some evidence of understanding the mathematical concepts and procedures in the NAEP content strands.<br>Fourth-grade students performing at this level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve simple real-world problems in all NEAP content areas. Students at this level should be able to use – though not always accurately – four-function calculators, rulers and geometric shapes. Their written responses are often minimal and presented without supporting information. |
| Proficient | This level represents solid academic performance for each grade assessed. Students reaching this level | Fourth-grade students performing at the proficient level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content strands.<br>Fourth-graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanation of how they were achieved. |
| Advanced | This level signifies superior performance | Fourth-grade students performing at the advanced level should apply integrated procedural knowledge and conceptual understanding to complex and non-routine real-world problem solving in the five NAEP strands.<br>Fourth graders performing at the advanced level should be able to solve complex and non-routine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution process by explaining why, as well as how, they were achieved. They should be beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely. |

Instead of using content standards, Mills and Jaeger (1998) used the actual test content to develop the descriptors. To do this, the researchers designed seven steps to be followed:

1. Convening and instructing the panel of subject matter experts;
2. Reviewing the items and exercises in the test booklet;
3. Reviewing the test framework;
4. Reviewing the generic (policy) definitions of the performance categories;
5. Linking of test content with the generic definitions;
6. Defining students' abilities associated with each performance category; and
7. Development of consensus descriptions of each of the performance categories.

When the researchers compared the performance descriptors produced by using content standards with those produced by using test content, they found out that they were different. When cut scores were set based on these two sets of descriptors, the student classification results were also different, suggesting that the descriptors have important effects on the cut scores. The major weakness of developing descriptors using test content is that new descriptors would have to be developed for each form of the test since the items will be different from form to form.

## 2.5 Misclassification Errors

It has already been pointed out that standard setting exercise provides information that can be used for making decisions. The information provided is derived from the test administered and/or examinees' test scores. Both of these can have measurement errors. For example, the process of test construction can introduce measurement error if the test

does not appropriately represent the syllabus. Examinees' test scores can also contribute to measurement error through, for example, lucky guesses on a multiple choice exam, scorers' mistakes, a level of test-wiseness that allows test-takers to detect correct responses even when they lack a command of the material, etc. Collectively, these can result in a higher or lower score than deserved. Thus the persons located immediately below or above the cut score are indistinguishable because there is error attached to their scores. Also, there is some error in the cut score too, and all these errors contribute to misclassification errors (Tanner, 1996).

Errors can also be introduced in the cut scores through particular selection of judges to participate in a standard setting study, the vargaries of applying a standard setting method, the qualifications of judges, or the quality of training.

There are two kinds of misclassification errors: *False positive* errors occur when one who is not competent is mistakenly classified to be so. *False negative* errors, on the other hand, occur when candidates who are competent are judged not to be. Because of these measurement errors, score distributions of the competent and the not competent can overlap, so that the lower competent examinees are in danger of being classified as not competent (false negative errors), and the most able of the not competent are fortunate to be misclassified as competent (false positive errors).

As tests are never error-free, these errors are inevitable in testing situations. However, the questions to be answered are: Which type of error is more serious? How can it be minimized?

Part of the difficulty in making competency decisions is that false positive and false negative classifications are inherently related. One can minimize errors in one

direction only at the expense of increasing errors in the other. False negatives can be reduced by lowering the cut score, but the result will be more false positives as more of those not competent are judged to be competent. On the other hand, false positives can be reduced by raising the cut score, but the inevitable result is a lot of false negatives. "Setting a sensible cutscore requires the determination of which type of error is more harmful" (Zieky, 2001, p. 46). The more threatening of the two misclassification errors depends upon the circumstances. Those judging the competency of prospective surgeons may be willing to accommodate a significant level of false negative errors so that the risk of qualifying an incompetent surgeon, a false positive error, is minimized. On the other hand, if there are inadequate numbers of professionals (e. g. teachers), the decision to accommodate relatively high levels of false positive errors may be tolerable.

## 2.6 Guidelines for Running a Standard Setting Study

One of the goals of a standard setting process is to ensure that the cut score is estimated using sound procedures, and that there is a firm basis for defending the selection of a particular cut score (Reckase, 2001). To ensure that a standard setting study produces defensible results, Hambleton (2001) outlined eleven steps to be followed when deriving cut scores using test-centered methods.

1. Choose a panel (large and representative of the stakeholders).
2. Choose one of the standard setting methods, and prepare training materials and finalize the meeting agenda.
3. Prepare descriptions of the performance categories (e. g. basic, proficient, and advanced).

4. Train panelists to use the method (including practice in providing ratings).
5. Compile item ratings and or other ratings data from the panelists (e.g. panelists specify expected performance of examinees at the borderline of the performance categories.
6. Conduct a panel discussion; consider actual performance data (e. g. item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on interpanelist and intrapanelists consistency.
7. Compile item ratings a second time that could be followed by more discussion, feedback, and so on.
8. Compile panelist ratings and obtain the performance standards.
9. Present consequences data to the panel (e.g. passing rate).
10. Revise, if necessary, and finalize the performance standards, and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards.
11. Compile validity evidence and technical documentation.

In the course of conducting a standard setting study, it is imperative to satisfy the "due process" requirement (Camilli, et al., 2001; Collins, 1995). The American National Institute (1993), defined due process, as quoted by Collins (1995), as follows:

> Due process means that any person (organization, company, government agency, individual, etc.) with a direct and material interest has a right to participate by: a) expressing a position and its basis, b) having that position considered, and c) appealing if adversely affected. Due process allows for equity and fair play. (p. 207)

According to Carson (2001), due process demands the state or state actors to treat individuals fairly, from a substantive, as well as from a procedural perspective

### 2.6.1 Evaluation of a Standard Setting Study

It has already been pointed out that the outcomes of a standard setting process can be used for making classification and certification decisions, among others. These decisions can have lasting impact on the people affected. But it has generally been established that different standard setting methods produce different cut scores (Andrew & Hecht, 1976, cited by Zieky, 2001) resulting in different classifications of examinees. This is in agreement with Jaeger's (1989) conclusions of the review of the literature on the comparability of standard setting methods: "Different standard-setting procedures generally produce markedly different standards when applied to the same test, either by the same judges, or by randomly parallel samples of judges" (p. 497). This is to be expected because different methods define minimal competency in different ways (Hambleton, 1978, cited by Zieky, 2001). For an examining agency, this is not necessarily an issue as long as the same technique is used year after year. However, it is still important for the measurement community to have a set of criteria that can operationally define the effectiveness of any single standard setting study.

Because of lack of agreement in standards set by different methods, some authors have written against standard setting itself. Shepard (1979, quoted by Zieky, 2001), advised people to "avoid setting standards whenever possible" (p.25). Glass (1978, quoted by Zieky, 2001) went further to say "setting performance standards on tests and exercises by known methods is a waste of time or worse" because all the methods were

"arbitrary" (p.24). But Popham (1978, quoted by Zieky (2001) counter-argued that, while performance standards were set judgmentally, it was incorrect to equate human judgment with arbitrariness in this negative sense. Mehrens and Cizek (2001) also defended standard setting saying: "To argue against standard setting is to, in effect, argue against making categorical decisions" (p. 479), which are unavoidable in education. This is echoed by Hambleton (1978) cited by Camilli et al. (2001), who asserted that instructional decisions cannot be made without cut scores. Whatever one's position regarding standard setting, the need for agreed guidelines is pertinent.

Indeed, there is a need for establishing formal criteria for evaluating a standard setting study. So what features should characterize a sound and defensible standard setting procedure? So far, different authors have provided guidelines with some variations from author to author. Berk (1986) proposed two criteria of defensibility – technical adequacy and practicability. The criterion of technical adequacy requires the standard setting method to: yield appropriate classification information; be sensitive to examinee performance; be sensitive to instruction or training; be statistically sound; identify true standard; and yield decision validity evidence. The practicability criterion demands that: the method be easy to implement; to compute; to interpret; and be credible to laypeople. For his part, Van der Linden (1995) formulated six criteria – explicitness, efficiency, unbiasedness, consistency, feasibility, and robustness – to be used to discriminate between better and worse standards.

Hambleton (2001) also formulated 20 questions that can be used to evaluate a standard setting study. Norcini (1997) proposed criteria of credibility and comparability. Under the criterion of credibility he talks about the quality of standard setters, how the

standards were set, and reasonableness of standards set. For the comparability criterion, he proposed that the content and performance of the test forms be the same, procedures for adjusting cut scores produce good results, and the results of the equating process be realistic. The evaluation of a standard setting study will also consider whether the "due process" requirement was taken into account when conducting it. From the definition given above, due process implies allowing people with direct or material interest to participate in the study. However, it is important to note that some authors are against the full observance of this requirement. Jaeger (1991) argued that standard setting exercises should involve subject matter experts, not policy makers. In his support, Norcini and Shea (1997) contended that standard setters should be leaders in their field, and it is not appropriate to ask non-experts to make judgments that require knowledge of content. In addition, Berk (1996) cited by Bennet (1998) expressed his view that a broad based panel of the most qualified and credible judges should be selected.

According to Kane (1994), validation of standard setting method involves consideration of policy and descriptive assumptions. The policy assumption claims that the performance standards are appropriate, given the purpose of the decision. The descriptive assumption claims that the cut scores correspond to a specified performance standard, in the sense that examinees with scores above the cut score are likely to meet the standard and examinees with scores below the cut score are unlikely to meet the standard. The process of validating a standard setting method basically involves evaluating these assumptions.

There are three types of evidence that can be used to evaluate the assumptions, and thus validating the standard setting method. These are procedural evidence, internal

consistency evidence, and evidence based on external criteria. Procedural evidence focuses on the appropriateness of the procedures and the quality of the implementation of these procedures. "Poor procedures or failure to implement procedures in an appropriate way can destroy our confidence in the resulting passing score and performance standard" (Kane, 1994, p. 437). Procedural evidence is a widely accepted basis for evaluating policy decisions. The legitimacy of a policy decisions is "evaluated in terms of general criteria, such as the reasonableness of the decision and the fairness and legitimacy of the procedures used to arrive at the decision" (Kane, 1994, p. 445). The legitimacy of the standards requires that the final judgment is not arbitrary or capricious (Reckase, 2001, p. 211). One of the most successful ways of demonstrating the rationality and reasonableness of passing standards is evidence of procedural validity (Plake, 1998, cited by Carson, 2001), and evidence of procedural validity focuses on who set the standards and how they did it (Carson, p. 431).

Evidence of internal consistency also provides support for the validity of a set of standards. Evidence of internal consistency includes: size of the standard error of the cut score and item-level data, such as proportion of examinees around the cut score answering an item correctly. This type of evidence is particularly relevant to the descriptive assumption, which posits a correspondence between the performance descriptors and the cut score. If the evidence does not support the relationship, it may be possible to correct the problem before the process is finalized.

Table 2.2 Summary of Criteria for Evaluating Standard-Setting Procedures

| Evaluation criterion | Description | Sources |
|---|---|---|
| Procedural | | |
| Explicitness | The degree to which the standard setting process was clearly and explicitly defined before implementation | Van der Linden (1995) |
| Practicability | The ease of implementation of the procedures and data analysis, and the degree to which procedures are credible and interpretable to laypeople. | Berk (1986) |
| Implementation of procedures | The degree to which the following procedures were systematic and thorough: selection and training of panelists, definition of the performance standards, and data collection. | Kane (1994, 2001) |
| Panelist feedback | The extent to which panelists feel comfortable with the process and with the cut score | Kane (1994, 2001) |
| Documentation | The extent to which features of the study are reviewed and documented for evaluation purposes | Cizek (1996b); Hambleton (1998); Mehrens (1995) |
| Interpanelists consistency | The consistency of item ratings and cut score across panelists; includes "caution indices," whereby panelists are flagged whose ratings are inconsistent with the majority. | Berk (1996); Cizek (1996b); Jaeger (1988, 1991) |

Table 2.2 Continued

| Evaluation criterion | Description | Sources |
|---|---|---|
| Internal Consistency within method | The precision of the estimate of the cut score, or the extent to which same cut score would be obtained if method were replicated | Cizek (1996b); Kane (1994, 2001); van der Linden (1995) |
| Intrapanelist consistency | The degree to which a panelist is able to provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds | Berk (1996); Cizek (1996b); van der Linden (1982) |
| Interpanelists consistency | The consistency of item ratings and cut score across panelists; includes "caution indices," whereby panelists are flagged whose ratings are inconsistent with the majority. | Berk (1996); Cizek (1996b); Jaeger (1988, 1991) |
| Other measures | The consistency of cut scores across item types, content areas, and cognitive processes | Kane (1995) |
| External Comparison to other standard setting methods | The consistency of cut scores across replications with other standard setting methods. | Kane (1994, 2001) |
| Comparison to other sources of information | The relationship between decisions made using the test to other criteria (e.g., grades, performance on a similar test, etc) | Berk (1996); Giraud et al. (2000); Kane (1994, 2001); Shepard et al. (1993) |
| Reasonableness of cut scores | The extent to which the resulting cut scores are feasible or realistic, including impact on pass rate | Kane (1998); van der Linden (1995) |

Note: From Pitoniak (2003). Reproduced with permission.

External evidence can also be used to validate standard setting results. By comparing standard setting results with other decisions, such as other assessment-based decisions, or results of other standard setting studies, it is possible to assess the appropriateness of the proposed cut score. Another way to demonstrate external evidence is to examine the impact of the cut score. The reasonableness of the results produced by the cut scores may determine their acceptability. All these types of evidence are summarized in Table 2.2.

## 2.7 Review of Some Standard Setting Studies

In this section, literature on studies that have a direct bearing on the present study is reviewed. The literature in this section is organized into: the role of training judges in standard setting, maintaining examining standards, standards set by different panels, and studies on MSCE standards.

### 2.7.1 The Role of Training Judges

Evidence from literature indicates that training standard setting judges can greatly minimize variability due to judges. Training will make them fully understand the process they are to follow and what is required of them (Berk, 1996; Kane 1998; Mills, 1995; Fehrmann et al., 1991, cited by Cresswell, 1996). Hambleton et al. (2000) provided evidence that judges are capable of making the necessary judgments when they have been properly trained. Raymond and Reid (2001) proposed three criteria that can be built into processes for determining whether a judge is well-trained: standard-setting ratings should be stable over time; standard-setting ratings should be consistent with the relative

difficulty of the items; and standard-setting ratings should reflect realistic expectations. Mills, Melican and Ahluwalia (1991) also supported the need to train judges to ensure they have a common understanding of minimal competence as it applies to a particular body of knowledge and skills.

> Without a common understanding of the process and a common definition of minimal competence, differences in item ratings may be more related to background variables of judges than to real differences in perceived item difficulty. (p. 7)

Thus, there is an established body of evidence that judges participating in a standard setting study need to be well trained for their task and must have a clear understanding of the work they are required to do. According to Rudner (1992), training judges aims to achieve three objectives: to familiarize the judges with the measures that they will be working with; to ensure that the judges understand the sequence of operations that they must perform; and to explain how the judges should interpret any normative data that they are given. Among other things, the content of the training will include: description of purpose of the examination, review of the examination development processes, a description of the various uses of examination results, a general overview of the standard setting methodology, concept of false positives and false negatives, and practicing estimating minimum performance level (Raymond and Reid, 2001). It is also necessary that the judges be given the opportunity to ask questions and discuss the process.

However, it is important to emphasize that for training to yield successful results, it is imperative that the selection of judges be done properly. This issue has been discussed extensively under "evaluation of a standard setting study".

### 2.7.2 Maintaining Examination Standards

For all exams, pass/fail decisions must be made, and such decisions need to be the same over time, and for all different forms of the test (Norcini & Shea, 1997). This is one of the major challenges facing examining institutions today: to ensure that standards remain the same over time. Examining institutions need to ensure that the cut score established each year represent the same level of proficiency in the subject: it should be just as hard to achieve the cut score this year as it was in the previous years. When this is achieved, then fairness between cohorts and comparability of inter-year grades will also have been achieved.

Apart from fairness and comparability of grades, maintenance of examination standards is necessary for the measurement of change or growth in students' attainment levels. From time to time educational reforms do take place and it becomes imperative to know how and to what extent such reforms have affected the outcomes of the system. In addition, examining agencies are sometimes required by their governments to provide external validation and monitoring services to ensure consistency of standards (Wolf, 1996). This becomes easy to do if examination standards are not changed. This makes true the psychometric saying that: "If you want to measure change, don't change the measure".

Over the years psychometricians have devised and used various methods for maintaining examination standards. One method has been to develop examinations of equivalent difficulty and maintain the same cut scores from one year to the next. But different forms of the examination are rarely equal in difficulty (Angoff, 1971). The

difficulty in developing examinations of equal difficulty is also echoed by Newton (1997):

> ... it is extremely complex for paper setters to gauge how hard candidates will find their questions, and a paper may be more or less difficult from one year to the next even though questions on similar topics are asked. (p. 229)

Hambleton (2000) reported another method that was used in Canada, where an assumption was made that successive cohorts of students were of equal quality. Based on this assumption a common passing rate would be established. This also, "does not allow consistency of standards across time because characteristics of cohorts taking the examination may change" (Newton, 1997; p. 229). In addition, this approach is against *standard 14.7*, which prohibits adjustment of cutoff score to regulate the proportion of people passing the test (AERA, APA, NCME, 1999).

Another approach would be to use the same test to successive cohorts of examinees and use the same cut score. The danger with this approach is that in the long run the items in the test will have different relevance with repeated administrations. Further, the repeated administrations poses a security risk. Some students may memorize the items or their content and reveal them to the next group of examinees. Thus for many reasons, including security, different forms of the test are used.

The most common way practitioners ensure comparability of standards is by test score equating. Test score equating is conducted to establish equivalence between test scores. An equating function f(x) is determined to map the raw scores obtained from a newer test form into raw scores obtained from an older test form. This means that some items are common to successive forms of the examination. These common items are used

to estimate the relative level of ability of candidates across forms of the examination. The actual cutoff score is adjusted based on candidates' performance on common items in order to maintain the standard. If equating has been successful, it is possible to compare students who take different forms of the test.

Where test items are disclosed following administration, as is the case in Malawi, it is not possible to do statistical equating of test scores. In such a situation, judgmental equating, also known as social moderation (Waltman, 1997) or linking, becomes very necessary. Some items from the previous years can be intertwined in the present years' examination and subject matter experts can be asked to rate all the items together. The ratings on the intertwined (common) items can then be used as the means of controlling the difficulty of new test forms so that they will be comparable to earlier test forms (at least judgmentally). If the ratings of the intertwined items are different from the way they were rated the previous year, then this year's ratings of all items can be adjusted to the scale of last year. Lorge and Kruglov (1952, 1953) cited by Thorndike (1982) applied the method and found that judges demonstrated moderately good agreement in appraising relative difficulty of test items, but differed widely in the absolute difficulty level that they assigned to the test items. Lorge and others also found that if the judges are provided with common items of known difficulty from a previous administration, their level of agreement in terms of absolute values of level of item difficulty improves.

Norcini (1990) compared the results of judgmental equating with those of test score equating. He assembled four test forms of approximately equal difficulty through random assignment of items to the forms. Each form had the same common items. Four groups of judges received initial briefing together, then they rated some items in their

separate groups. They completed their rating of the remaining items individually. Comparison of equating results of the cut scores indicated that judgmental linking produced more accurate results than statistical linking, especially when the cut score was relatively extreme and examinee samples were small. In the case of MSCE, there are usually very small numbers in the distinction category. Norcini's findings would suggest that judgmental linking would be suitable for MSCE, at least for the distinction category, because there are usually very small numbers.

Stobart et al (1990) also employed judgments of raters to ascertain whether the Geography grades awarded by the six examining groups in the United Kingdom were comparable in terms of level of attainment which they represented. Each examining group based its standards on the national criteria, which provided a common framework for all the groups. The judges, who came from each of the six examining groups, made holistic ratings of the actual borderline candidates' work at three grade boundaries. They did not rate the work of the candidates who took their own examination, but their ratings were based on the standards applied by their own examining groups. The results showed a broad equivalence between the examining groups' standards. However, there were differences at specific grade boundaries. No group was consistently rated lenient or severe across all the three grade boundaries of A/B, C/D, and F/G. The differences at specific grade boundaries suggested that the Examining Groups interpreted the national criteria somewhat differently.

In order to benefit from the advantages of the various methods of standard setting, and to ensure maintenance of standards, Whetton, Twist, and Sainsbury (2000) utilized four different methods in their study: statistical equating to the previous year's test;

equating to an anchor test; Angoff-type; and script scrutiny. The equating methods were based on empirical data while the last two were based on expert judgments. Cut scores were set during a meeting between the test development agency and the responsible government agency. To arrive at a cut score, the meeting considered evidence from all the four methods.

Wheton et al (2000) observed that triangulation of methods had the potential to improve the acceptability and quality of standard setting results, since it captures the advantages of all the methods used while avoiding the disadvantages of the individual methods. However, triangulation of methods creates its own disadvantage: how to combine information from several sources to arrive at a single decision on the cut score. The study recommended "formal weighting for the four types of evidence…which would allow the arithmetic procedures to be used for their combination, rather than private individual judgements" (p. 16).

### 2.7.3 Standards Set by Different Panels

As it has already been pointed out, one problem with standard setting studies is that different standard setting methods produce different results. The situation is further worsened by the finding that results of standard setting also depend on the set of judges that participate in the study. Results of a study conducted by Jaeger, Cole, Irwin and Pratto (1980) confirm that different judges will produce different results. Jaeger and his colleagues had three panels consisting of samples of teachers, administrators and counselors, respectively. They independently set passing standards on one of the North Carolina school achievement tests. There were wide variations in the standards set. On

the reading test, the proportion who would have failed ranged from a low of 9% to a high of 30%. The situation was worse in mathematics where failure rate ranged from a low of 14.4% to a high of 71.1%.

In their study, Good and Cresswell (1988) cited by Cresswell (1996) found that the percentage of candidates whose subject grade changed if one awarding team was substituted for another was 13% in French, 17% in Physics, and 38% in History. However, it is not reported whether the teams received the same training or discussed their results. In general, different groups of judges will set different standards when using the same method, especially if the judges represent different interest groups.

However, Kingston, Kahl, Sweeney and Bay (2001) found consistent results across panels. In their study, they involved three states: Maine, Massachusetts, and Wyoming. They implemented a Body of Work (BoW) method in which each state had its own panel. The three states used the same performance level descriptors, although the states had different names for performance levels. All the states had four performance levels. The BoW method produced about the same percentage of students in the highest and lowest performance levels. Percentages in the middle performance levels were different.

In another study, Plake, Impara, & Irwin (2000) examined the intra- and inter-rater consistency of item performance estimates using the Angoff method. Their study found that item performance estimates were consistent within and across panels and across years.

## 2.8 Studies on MSCE Standards

The literature search located three studies that investigated MSCE standards. These studies are presented in this section.

### 2.8.1 AEB/MCETB Comparability Study

This was the first study to be conducted on MSCE standards. It was conducted in 1979, when the examination was known as the Malawi Certificate Examination (MCE). The study compared the MCE standards with those of General Certificate of Education (GCE) administered by the Associated Examining Board (AEB). From the outset it was agreed between the two boards, that is, the AEB and the Malawi Certificate of Education and Testing Board (MCETB), that grades 1-6 (distinction and credit grades) of the MCE would be equivalent to grades A to C of the GCE (MCETB, 1979). In other words, a GCE pass performance (Grade C) would be equivalent to the lowest MCE credit performance (Grade 6).

A cross-moderation technique, which involved Chief Examiners from each Board reviewing the standards of the examinations of the other Board, was used. The Chief Examiners studied the syllabuses, question papers, and scoring schemes of the other Board. They also studied 100 answer booklets of examinees from the other Board. The scripts were selected in such a way as to represent the entire score range of the examination. The Chief Examiners were asked, with their own Board's standards in mind, to record where they would have placed their grade boundaries within the array of scripts they were given. They had no idea as to where the other Board had placed its own grade boundaries. By collecting such judgments from the examiners of both Boards it

was hoped that it would be possible to gain an overall view of the degree of comparability that existed between the grading standards of the two Boards. The findings of the study indicated a reasonably high degree of equivalence between the AEB O-level grades A-C and the MSCE grades 1-6.

### 2.8.2 Trends of Performance at Credit and Distinction Levels

This study investigated the pattern of performance of each subject at credit and distinction levels from the time MCE examination was introduced in 1972 to 1980. Proportions of examinees obtaining credit and distinction grades for each subject each year were computed and compared. It was observed that the standard of performance varied from year to year in almost all subjects (MCETB, 1980). The Presidential Commission of Inquiry into MSCE results (Malunga, 2000) made similar observations.

### 2.8.3 Application of Standard Setting Methods in Public Examinations

Zoani (1989) used the 1988 MSCE Physical Science examinees' answer booklets to compare standards set by five standard setting methods. The five methods were: the norms approach, the Angoff, the Hofstee, the borderline, and the contrasting-groups methods. In the norms approach the judges were required to assume that examination papers for different years were comparable in terms of difficulty, the examinee population for different years were equivalent in terms of ability, and the same content was examined in each administration. The researcher also gave the judges a definition of the minimally competent examinee, which they were required to use when making judgments.

Among other findings the study observed that the norms approach produced results similar to those of the Board. This finding should not be surprising because the method made similar assumptions as those used by the Board. The study also found out that in spite of using a common definition of the minimally competent examinee the judges appeared to have understood the definition differently, judging by their cutoff scores:

> The results seem to suggest that different judges will have different conceptualization of what constitutes the minimally competent examinee even when the definition for the minimally competent examinee has been given. (Zoani, 1989, p. 463)

This finding is also not surprising, because the definition was not accompanied by performance descriptions, which serve as common frameworks for the judges. A study by Stobart et al (1990) also obtained a similar finding where different examining groups in the United Kingdom differed in the way they classified examinees, in spite of using the same national criteria.

Further, Zoani's study did not ask the judges to work on sample items before making judgments on the main test. If this were done it would have been possible to check whether the judges had a common understanding of the process and the definition. The need for a common understanding is also echoed by Mills et al (1991) who believe that without common understanding of the process and definition of minimal competence, differences in item ratings may be due to background variables of the judges and not real differences in perceived item difficulty (p. 7).

In the borderline and contrasting-groups approaches, the teachers used their experience and knowledge of what they thought were the capabilities of borderline

examinees to estimate their probabilities of correctly answering an item. Because their experiences were varied, it was difficult to achieve consistency across teachers.

## 2.9 Summary

This comprehensive literature review has identified several points of significance to the goals and design of the study. First, the importance of maintaining validated performance standards in credentialing and achievement examinations, of which MSCE is one, cannot be overstated. Psychometricians have used various methods for establishing and validating standards and many of these were described in the chapter. Where test score equating is not possible or not appropriate, Norcini (1990) has demonstrated that judgmental linking can produce even more accurate results than equating tests through test scores, and a number of studies (Stobert et al., 1990; Whetton et al., 2000) have successfully used the technique. These points impact directly on the approaches taken in this study

Second, the literature has also emphasized the need to define performance standards in terms of students' behaviors. This has the advantage of improving understanding of test results and interpreting test scores, besides describing what competences a given score represents (Haertel, 2002). In this study, performance standards were determined using performance level descriptors.

Third, it has also been demonstrated in the literature review that training of participants has the desirable effects of facilitating their understanding of the process and minimizing variability due to judges. This is one of the variables that this study investigated.

Finally, the importance of validation of performance standards comes through. When standards have been set, it is not enough to feel satisfied that the task is complete. It is necessary that the appropriateness of the performance standards be evaluated (see, for example, Hambleton, 2001; Kane, 1994, 2001; *Standards*, 1999). This study has used Kane's (1994) evaluation framework, which requires procedural, internal, and external evidence of validity.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

This chapter describes how the research problem was investigated. The list of questions to be answered is given first. This is followed by the design of the study. The standard setting method is given next. In the final section, the specific activities carried out to answer the questions are given.

## 3.2 The Research Questions

The following questions were addressed in this study:

1. What skills should students demonstrate in order to be graded, pass, credit, or distinction?
2. How would the standard setting results of two sub-panels using the same performance level descriptors compare?
3. Does the application of the same performance level descriptors yield consistent results over years?
4. How do the equated cut scores that are based on common items compare with those that are based on common judges?
5. How do the SMEs' ratings before and after scoring students' work compare?
6. How do the standards set by trained SMEs compare with those set by untrained SMEs using the same performance level descriptors?

## 3.3 The Design

All together, there were five panels. Twenty judges were trained together before being split into three panels. Two other panels were created, one from among the trained judges, and the other from untrained judges. All these panels are now described.

Panels 1: This panel consisted of seven trained judges. Five of the judges were males and two were females. Three were diploma holders, two were first-degree holders and two had masters. The panel rated the 2003 items.

Panel 2: This panel consisted of seven trained judges. The panel composition was identical to that of Panel 1. Like Panel 1, Panel 2 rated the 2003 items. The comparison of cut scores set by Panel 1 with those set by Panel 2 provided basis for answering Question 2.

Panel 3: This panel consisted of six trained judges. Four of the judges were males and two were females. Two judges in this panel were diploma holders, three were first-degree holders and one had masters. This Panel rated the 2002 items. The comparison of cut scores set by Panel 3 with the average of those set by Panels 1 and 2 allowed Question 3 to be answered.

Panel 4: This panel consisted of six trained judges. Three of the judges in this panel came from Panel 3 (who served as common judges). Of these three, two were males and one was a female. Two judges - one male and one female - came from Panel 1, and one female judge came from Panel 2. This panel rated 2003 examination items with some 2002 common items intertwined in the examination. The judges in this panel participated in scoring 2003 students' answers. The rating of items took place after the judges had participated in scoring students' answers.

The comparison of equated cut scores generated from common items (items were common to judges only, not students) with those generated from common judges allowed Question 4 to be answered. Further, a comparison of cut scores set by Panel 4 with the average of those set by Panels 1 and 2 allowed Question 5 to be answered.

Panel 5: This panel consisted of four untrained judges – three males and one female. Like Panel 4, judges in Panel 5 rated the 2003 examination items with some 2002 common items intertwined in the examination. The judges in this panel participated in scoring students' 2003 answers. The rating took place after scoring students' answers. The comparison of cut scores set by Panel 4 and Panel 5 answered the question of whether training has any impact on the cut score.

P-values and consequence data for the 2003 examination were not available since the study was conducted before the release of the 2003 results. However, 2002 item p-values (see Appendix F) and consequence data were available. Table 3.1 summarizes the experimental design.

Table 3.1 The Experimental Design

| **Session** | **Trained Judges** | | **Untrained Judges** | |
|---|---|---|---|---|
| | 2003 | 2002 | 2003 | 2002 |
| Before Scoring | Panel 1 (7)<br>Panel 2 (7) | Panel 3 (6) | | |
| After Scoring | Panel 4 (with 2002 common items and common judges) (6) | | Panel 5 (with 2002 common items) (4) | |

Note: Numbers of judges in each group are shown in parentheses.

## 3.4 The Method

An *item score string estimation* method (Loomis & Bourque, 2001) was used in this study. What follows are the details of how the method was applied.

### 1. Choosing judges

The participants were chosen based on their expertise and experience of teaching the subject at MSCE level. All twenty participants who were invited to participate in the study turned up. The researcher personally knew some of the judges, others were recommended by colleagues and Chief Examiners of the subject. All the judges were teachers or had been teachers in the subject at MSCE level. To ensure against bias of any kind the judges were drawn from a variety of schools – Public, Grant-Aided, and Private schools - from all the three regions of the country. Care was also taken to achieve gender balance. Apart from schools, the following institutions were represented:

MANEB;

Mathematics Moderation Committee;

Curriculum Developers – Malawi Institute of Education (MIE);

Mathematics Syllabus Committee;

Ministry of Education;

University of Malawi;

Domasi College of Education.

Judges were first contacted by telephone. Details of the training program, together with some training materials, were mailed to those who accepted to participate two weeks

ahead of the program (see Appendix G for invitation letter). Training materials included test papers, MSCE labels and policy definitions of performance categories (see Appendix H), and a tentative agenda (Appendix I). Judges were also advised to familiarize themselves with the 2002 and 2003 MSCE Mathematics questions before going to the workshop venue. This was meant to avoid spending too much time trying to understand test materials at the workshop. Other training materials, such as item p-values, scoring schemes, students' answer booklets, and 2002 Mathematics score distributions, were provided to judges at the workshop.

Upon arrival at the workshop venue, judges were each given a folder containing some training materials such as item p-values, scoring scheme, students' answer booklets, etc. They also completed a registration form (see Appendix J) from which their demographic information was obtained. The information included gender, teaching experience, qualification, age, and present involvement in MSCE Mathematics. Their ages and teaching experience ranged from 31-54 and 3-29 years, respectively. Table 3.2 shows how the judges were allocated to the panels.

Table 3.2 Distribution of Judges to the Panels

| Panelist characteristic | Panel 1 | Panel 2 | Panel 3 |
|---|---|---|---|
| Male | 5 | 5 | 4 |
| Female | 2 | 2 | 2 |
| Diploma Holder | 3 | 3 | 2 |
| First Degree | 2 | 2 | 3 |
| Masters | 2 | 2 | 1 |

2. Choosing a standard setting method

Because MSCE Mathematics uses only performance assessments with polytomous items, an item score string estimation (ISSE) method, which handles both dichotomously and polytomously scored items, was used. Another reason for choosing the method was that, like other methods which base their judgments on total test score, it is a compensatory model: it allows students to compensate for low performance on some exercises or tasks by achieving higher scores on other exercises or tasks. The fact that MSCE examinees are graded based on their total scores means that the process is compensatory, and the use of a compensatory model is appropriate. The method is also easy for judges to understand and to use (Impara & Plake, 1998).

In its original form, the method requires judges to estimate whether borderline examinees would correctly answer a dichotomously scored item. The method requires judges to give a *yes* or *no* answer if the borderline students can answer the dichotomously scored item correctly or incorrectly, respectively. The method can also be used for polytomously scored items by requiring judges to estimate the score borderline students would earn on each item. For a 3-point item, for example, judges would estimate whether a borderline examinee would score 1, 2, or 3 points.

In this study, the judges used performance level descriptors to determine item score points likely to be obtained by the borderline examinee for pass, credit and distinction categories. Judges were warned against confusing between typical performance of a category and minimum performance level for a category, that is, performance of a borderline examinee for a particular category. It was emphasized during

training that the judges' task was to estimate borderline performance, which is the lower boundary of each achievement category.

3. Preparing descriptions of the performance categories

The description of the quality of performance for each performance category is a crucial component of the standard setting process. Where the cut score is placed on the score scale, will very much depend on the clarity of the category descriptions. For the descriptors to be helpful in the standard setting process, they needed to be developed diligently and written clearly. One of the important requirements for the development of useful descriptors is the policy definition of the performance category (Appendix H). As the policy definitions are general in nature, that is, they are not subject-specific, the judges' task would be to translate these definitions into detailed performance level descriptors (see Appendix K) for their subject area, which, in this study, was Mathematics.

The workshop itself began with participants filling in their registration form (Appendix J). This was followed by self-introductions. Then the facilitator offered welcoming remarks, after which he delivered a comprehensive power point presentation covering important standard setting issues which included: purpose of the examination, development of the examination, processing of examination results, methods of maintaining examination standards, development of performance level descriptors, the standard setting method to be used, and introduced the topic of misclassification errors. Then the judges practiced performing item ratings on four sample items and then discussed their ratings. After training together, the judges were split into three panels described previously.

Since the process of developing performance level descriptors requires knowledge of the labels for performance categories and their definitions, this study used the same labels currently used by MANEB for MSCE grade categories. These grade categories, in ascending order, are fail, pass, credit, and distinction. To write the descriptions, the panelists used MANEB's policy definitions of the grade categories. As these definitions were initially not available, the study first engaged MANEB officers responsible for policy matters to formulate these definitions. The officers were presented with examples of policy definitions from other examining agencies to help them understand the task and consider whether to adapt or adopt them. The document containing these policy definitions was among the training materials that were mailed to the participants ahead of the standard setting workshop.

At the workshop the judges first discussed these definitions before using them to derive the performance level descriptors. Since the policy definitions were not subject-specific, the judges were asked to translate these definitions into detailed performance level descriptors for MSCE Mathematics. They were told that the performance level descriptors must provide a direct link between MANEB's policy definitions of the levels of achievement and MSCE Mathematics content. In developing these descriptors, the judges were reminded that the achievement levels were cumulative in nature, that is, examinees in the higher performance category would surpass the requirements for the lower categories. Some examples of performance level descriptions from other testing agencies were presented to the judges.

Since the MSCE Mathematics syllabus already had the descriptors in the form of objectives to be mastered by students, the judges' task was simply to classify these

objectives into pass, credit or distinction categories, depending on the degree of difficulty of the objectives (see Appendix K). As will be shown later, the judges had to adjust the descriptors twice to ensure that they produced reasonable standards.

4. Train judges to use the method

Setting standards is a difficult, judgmental task. For good results, it is extremely important that the judges employed in the process are not only knowledgeable, but also well trained in the method. They need to fully understand the process they are to follow and what is required of them. Among other things, the judges need to be familiar with the measures that they will be working with, and understand the sequence of operations that they must perform. In this study, the item string estimation method of standard setting was presented to the participants. Using the performance level descriptors, the judges were requested to determine cut scores for four sample items taken from the 2000 MSCE examination papers. They first solved the problems. Then, the solutions and scoring guides for the four problems were provided. Considering the skills needed to solve the problems, the judges determined whether the borderline examinees would be able to perform the skills. They awarded a point for a skill they believed a borderline examinee would be able to perform, and zero if not. For each point awarded, they determined the performance level the skill belonged by simply checking its location on the performance level descriptor form. For each item, the points awarded to skills belonging to the same performance category were summed to get the item cut score for that category. All the item cut scores for each category were summed up to get the paper cut score for the category

The item cut scores set by individual judges were displayed on a chart. Judges with discrepant results were requested to explain the basis of their results. After some discussion, the judges reached a consensus as to which cut scores were appropriate.

As a way of facilitating the training process, the following training materials were provided: examination papers (Appendices A to D), scoring schemes, item p-values for the 2002 question papers (Appendix E), students' written answers, item rating forms (Appendices L & M), copies of the syllabus, and 2002 score frequency distributions and other descriptive statistics. The item p-values for the 2003 items were not available because the standard setting study took place before the scoring of 2003 examination.

5. Compile item ratings

After the judges had worked on the practice items, and when they had conceptualized the borderline performance standards for the categories, they translated them into operational standards, that is, the cut scores. They did this by estimating the number of score points a borderline examinee would obtain on each item by considering the skills on the solution process a borderline examinee would be able to demonstrate. The points awarded to the skills belonging to the same performance category were summed up to obtain the item cut score for the category.

As described in Table 3.1, the judges were split into three panels after they had trained together. Panels 1 and 2 were requested to set cut scores on the 2003 MSCE Mathematics papers while Panel 3 set cut scores on the 2002 papers. Before setting cut scores on the items, the judges were asked to individually solve problems on Paper 1. The aim of the exercise was for judges to appreciate what was required to arrive at the answers, and also for them to become familiar with the tasks. It must be mentioned that in

the invitation letter (see Appendix G), the judges were advised to familiarize themselves with the questions on the 2002 and 2003 MSCE Mathematics question papers. Judging by the speed with which they finished solving the problems for Paper 1, it was clear that the judges had heeded the advice.

6. Conduct panel discussion

After the participants had individually rated the items, they were given an opportunity to consider and discuss each other's explanations and justifications for their decisions. Judges with different ratings from the others were encouraged to explain the basis for their ratings. The panels were advised to compute the mean in case of unresolved differences. All the panels reached consensus, without resorting to computing the mean.

7. Compile item ratings a second time

When panel results were reported, it was observed that the initial standards were set too high: the 2002 consequence data showed an unacceptable level of failure rate. As a result of this, the judges reclassified the performance level descriptors. Using the new classification of performance level descriptors, the judges set new cut scores. When these were presented, it was observed that a few performance descriptors needed to be moved upwards again, because the new cut scores were considered too low. When this was done, the judges expressed satisfaction with the final classification of performance level descriptors, and set cut scores based on the new classification of performance level descriptors

8. Compile judges' ratings and obtain the performance standards

After the judges made their final item ratings they discussed them in their separate panels and reached a panel consensus. For each category, they summed the item cut scores to obtain the overall performance standard for the category. All three panels presented their results for the three performance categories.

9. Present consequence data to the panel

After obtaining the performance standards, consequence data for only 2002 examination in the form of proportions of examinees falling in each performance category was provided. The judges expressed satisfaction with the final outcome.

10. Revise if necessary, and finalize the performance standards, and conduct judges' evaluation of the process itself and their level of confidence in the resulting standards

When the final ratings were submitted, evaluation of the whole standard setting process followed. The purpose of the evaluation exercise was to gather information from the judges about their level of satisfaction with the performance descriptors, the training, the standard setting process, and the final standards. These pieces of information provided evidence for establishing the validity of the performance standards.

11. Compile validity evidence and technical documentation.

The whole standard setting process, from choosing judges to the final results, including evaluative results was documented. The documentation serves as the needed support for the validity of interpretations made from scores on the test for which standards were set (Pitoniak, 2003).

## 3.5 How the Research Questions were Answered

This section describes the analyses that were intended to answer the research questions. The analyses for each question are described separately.

### 1. What skills should students demonstrate in order to be graded pass, credit, or distinction?

To define the skills for the performance categories of pass, credit, and distinction, the judges first discussed the meanings of MANEB's policy definitions of the performance categories (Appendix H). Since these definitions were initially not available, MANEB was requested to formulate them. Examples of policy definitions from other examining agencies were presented to the officers who were assigned the task.

To derive the descriptors, the judges considered the meanings of the policy definitions. As the policy definitions were general in nature, that is, they were not subject-specific, the judges' task was to translate these definitions into detailed performance level descriptors for the subject area of Mathematics. The judges were told that the performance level descriptors they were going to develop should provide a direct link between MANEB's policy definitions of the levels of achievement and MSCE Mathematics content. They were also told that the descriptors would constitute competences that the examinees needed to demonstrate in order to be classified in a particular grade category. In developing these descriptors, the judges were reminded that the achievement levels were cumulative in nature, that is, examinees in the higher categories would have surpassed the requirements for the lower categories.

Since the MSCE Mathematics syllabus already had the descriptors in the form of objectives to be mastered by the students, the judges' task was simply to classify these

objectives into appropriate performance categories. To ensure that the descriptors were indeed appropriate for the categories, judges considered the reasonableness of the cut scores that were generated from the descriptors. Twice the descriptors were adjusted because the impact of the resulting cut scores did not look reasonable.

2. How would the standard setting results of two panels using the same performance level descriptors compare?

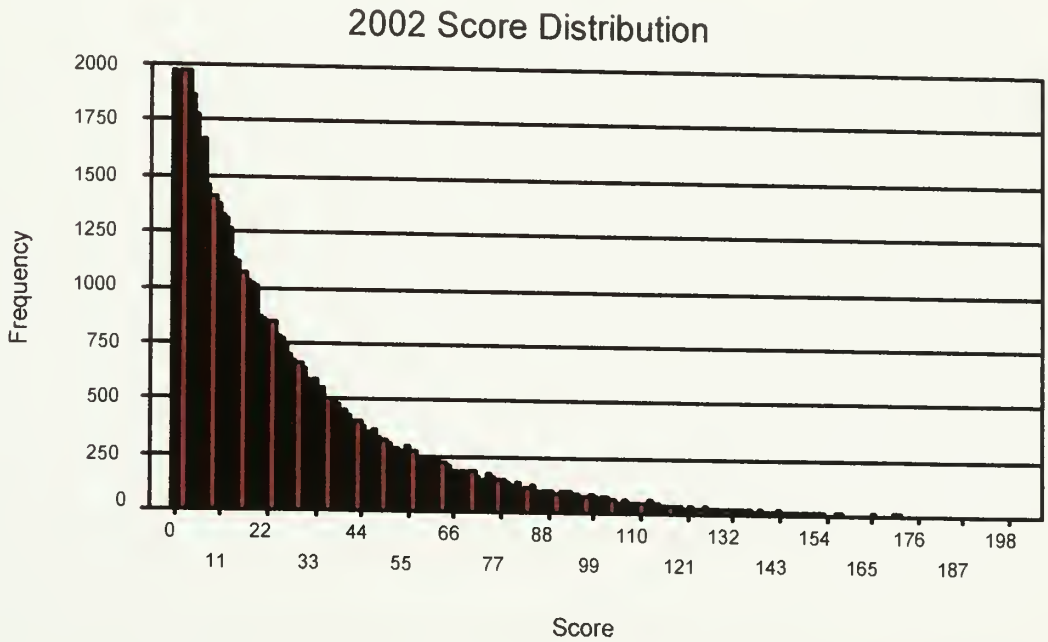
Panel 1 and Panel 2 were requested to set cut scores on the 2003 Mathematics papers. The judges studied the solution process for each item. They individually determined the skills (performance level descriptors) involved in the solution process. They also determined the performance categories to which the skills belonged. They grouped together all the skills belonging to the same performance category. Then they added up all the points that had been awarded to the skills belonging to the same performance category. This was the item cut score for that category. They did the same for the other performance categories. The judges repeated the process for all the items on the test. They discussed their individual item cut scores, and an opportunity for them to adjust their cut scores was given. When they were satisfied with their item cut scores, they added up all the item cut scores for each category to obtain the test cut score for that category. Cut score differences for each performance standard were computed. Correlations of the item ratings by the two panels at each performance standard were also computed to determine level of agreement of their item ratings.

3. Does the application of the same performance level descriptors yield consistent results over years?

To determine if application of the same performance level descriptors yielded comparable results on different forms of the examination, another panel, Panel 3, was

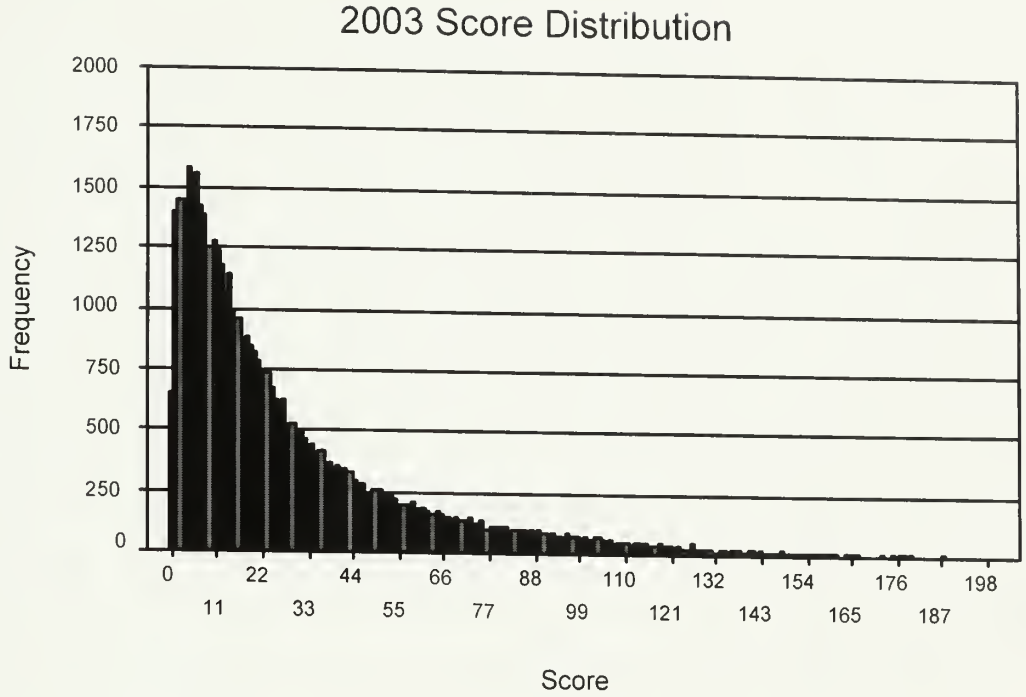
requested to set cutoff scores on the two papers of 2002 MSCE Mathematics. To do this, the judges followed the same process followed by Panels 1 and 2. Panel 3 results were compared with the average of Panels 1 and 2 who set cut scores on the 2003 examination. The comparison involved computing differences between corresponding cut scores and proportions of examinees falling in each performance category. Figures 3.1 and 3.2 show how the students' scores were distributed.

Figure 3.1




Note: Mean = 27.63; Standard deviation = 28.23; Minimum score = 0; Maximum score = 196; Number of examinees tested = 55997

Figure 3.2




Note: Mean = 29.06; Standard deviation = 30.16; Minimum score = 0; Maximum score = 198; Number of examinees tested = 46466

In making the comparison, an assumption that the two cohorts were of equivalent ability was made. This assumption makes sense because, generally, changes in any two successive years are expected to be imperceptible. Over a long period, of course, changes in the quality of students are inevitable. To ensure that this assumption was not violated, ten well-established and stable schools were sampled and the standard setting results were applied to the work of these schools. The reason for choosing stable schools was that successive cohorts of students from these schools were assumed to have reached about the same level of attainment by the time they completed their secondary education. Large differences in the proportions of students falling in the various grade categories may be attributed to the differences in standards set by the two panels. Comparison of the

cut scores for the two years was also made on the results for the whole population of examinees.

4. How would the equated cut scores that are based on common items compare with those that are based on common judges?

Comparison of the 2002 and 2003 cut scores which were set by different panels can be problematic because the tests may not necessarily be of the same difficulty. If, for example, the 2003 test is a bit easier, a panel might set higher cut scores. Without the knowledge that the 2003 test was easier, it would simply appear that the 2003 panel was more lenient in its judgment.

One solution might be to equate the 2003 test scores to the 2002 test scale by using some common items to the two tests in the judgmental process or some common judges. So the 2003 cut scores were equated to the 2002 scores via common items and common judges, and then the cut scores could be compared.

To obtain the equated cut scores that were based on common items reviewed by the 2002 and 2003 panels, ten of the 48 items from the 2002 examination were intertwined in the 2003 examination. (The common items had a total of 53 points (marks)). Panel 4 was asked to rate the items. This was done immediately after scoring the 2003 examinees' answers. The judges were told that the purpose of doing the exercise a second time was to see if they would maintain the standards they used during the first ratings. The 2003 ratings were adjusted to the 2002 scale based on how the 2002 items were rated relative to the first rating. To do this, the linear equating method was used.

Similarly, to generate the equated cut scores that were based on the judges who were common to both the 2002 and 2003 panels, the ratings by the three common judges (three judges who rated the 2002 items during standard setting workshop) in Panel 4 were

used to adjust the 2003 ratings to the scale of 2002. Because Panels 1, 2, and 3 had reached consensus, the linear equating formula could not be used for this part of the study, because the formula requires a standard deviation. Thus, instead of linear equating, mean equating was used. The two sets of equated cut scores were compared by computing their cut score differences.

To guard against practice effects, the common judges were not told in advance that they would be required to do the exercise again. This was to prevent them from deliberately remembering how they rated the items the first time. A period of four weeks was allowed between first ratings and second ratings to further minimize memory effects.

5. How do the SME's ratings before and after scoring students' work compare?

The ratings by Panel 4 were made after the judges had participated in the scoring exercise. Their ratings were compared with those that were made before scoring, that is, during the standard setting workshop. The comparison involved computing the cut score difference, the proportion of examinees in each performance category, and the correlations of item ratings at each performance standard.

6. How do the standards set by trained SMEs compare with those set by untrained SMEs, but using the same performance level descriptors?

Panel 4 and Panel 5 set cut scores using the same performance level descriptors that were developed during the standard setting workshop. As described already, Panel 4 consisted of trained participants, while Panel 5 consisted of untrained participants. The untrained participants were briefed about what the exercise entailed. They were also told that the purpose of the exercise was to assess the adequacy of the performance level descriptors in guiding the item rating process. After they understood what to do, they were given the rating forms, examination question papers, and scoring guide. There was

no practice exercise. There was also no opportunity for discussing their individual results. The means of item ratings for each panel were computed and added up to get the cut scores for the categories. Comparison of the cut scores set by the two panels was made by computing differences between average item cutoff scores and running a correlational analysis to see the level of agreement. Standard deviations of the cut scores for each panel were also computed to compare the degree of variability within the panels.

# CHAPTER 4

# PRESENTATION OF RESULTS

## 4.1 Introduction

In this chapter, the research results are presented. The results are presented for each research question separately. The first section presents competences the participants to the standard setting study thought were necessary for examinees to demonstrate in order to be classified in a particular performance category. The second section compares cut scores set by two different standard setting panels using the same performance level descriptors. This is followed by presentation of evidence that demonstrates whether the application of the same performance level descriptors on two forms of the examination can result in cut scores that represent the same level of proficiency. The fourth section compares results of equated cut scores based on common items with those that are based on common judges. The fifth section compares cut scores set before and after scoring examinees' answers. Following this is a section that compares cut scores set by trained subject matter experts (SMEs) and those set by untrained SMEs. The final section presents results from a survey given to judges in the study.

## 4.2 Competences Necessary for Classification in a Performance Category

One of the tasks the judges were requested to do was to develop performance level descriptions for the various grade categories. As explained in chapter 3, these descriptions were already available in the Mathematics syllabus in the form of objectives to be mastered by students. The judges simply classified them in the various performance categories, depending on their perceived difficulty. For each performance category, the

judges identified the objectives that they thought a borderline candidate should master. Appendix F presents the classification of these descriptors.

It must be mentioned that the initial classification of these descriptors produced very high cut scores. For example, the 2002 pass score was set at 72 (out of a possible 200), which only 8.7% of candidates could reach, using the impact data. Some descriptors had to be moved to higher performance categories, and the resulting classification produced a pass cut score of 32 (of 200 points, a 16% of the test score points), which allowed 31.8% of the candidates to pass. Although the pass rate was reasonable, the participants were concerned that the pass score was too low, especially considering that some of the questions on the test had been taken from JCE work. Another adjustment was, therefore, performed that produced cut scores the participants were happy with. Table 4.1 compares the impact of the resultant cut scores following each adjustment. The detailed item ratings before and after each adjustment are presented in Tables 4.2 and 4.3.

Table 4.1 Impact of Cut Scores Set Before and After First and Second Adjustments

| | | Pass and above | Credit and above | Distinction |
|---|---|---|---|---|
| First rating | Cut score | 72 | 135 | 167 |
| | % Examinees | 8.69 | 0.67 | 0.09 |
| After first adjustment | Cut score | 32 | 102 | 150 |
| | % Examinees | 31.78 | 2.93 | 0.28 |
| After final adjustment | Cut score | 42 | 107 | 145 |
| | % Examinees | 22.71 | 2.38 | 0.39 |

Table 4.2 How the 2002 Paper 1 Cut Scores Changed After Adjustment of Descriptors

| | First Ratings | | | Second Ratings | | | Final Ratings | | | Item Max. Score |
|---|---|---|---|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Dist. | Pass | Credit | Dist. | Pass | Credit | Dist. | |
| 1 | 4 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 |
| 2 | 4 | 0 | 0 | 0 | 4 | 0 | 4 | 0 | 0 | 4 |
| 3 | 0 | 3 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 3 |
| 4 | 0 | 0 | 6 | 0 | 0 | 2 | 0 | 0 | 0 | 6 |
| 5 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 6 | 1 | 2 | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 7 | 1 | 5 | 0 | 0 | 6 | 0 | 0 | 6 | 0 | 6 |
| 8 | 0 | 4 | 0 | 0 | 2 | 2 | 0 | 4 | 0 | 4 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 10 | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 3 | 3 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 12 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| 13 | 4 | 0 | 0 | 2 | 2 | 0 | 0 | 4 | 0 | 4 |
| 14 | 5 | 0 | 0 | 0 | 2 | 3 | 2 | 3 | 0 | 5 |
| 15 | 4 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 2 | 4 |
| 16 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 3 |
| 17 | 1 | 4 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 5 |
| 18 | 0 | 0 | 4 | 4 | 0 | 0 | 1 | 0 | 3 | 4 |
| 19 | 5 | 0 | 0 | 2 | 3 | 0 | 0 | 5 | 0 | 5 |
| 20 | 0 | 4 | 0 | 0 | 0 | 4 | 0 | 4 | 0 | 4 |
| 21 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 3 |
| 22 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 23 | 6 | 0 | 0 | 2 | 4 | 0 | 6 | 0 | 0 | 6 |
| 24 | 0 | 4 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| Total | 41 | 33 | 17 | 17 | 38 | 24 | 25 | 38 | 19 | 100 |
| Paper cut score | 41 | 74 | 91 | 17 | 55 | 79 | 25 | 63 | 82 | |

Table 4.3 How the 2002 Paper 2 Cut Scores Changed After Adjustment of Descriptors

| | First Ratings | | | Second Ratings | | | Final Ratings | | | Item max. Score |
|---|---|---|---|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Dist | Pass | Credit | Dist | Pass | Credit | Dist | |
| 1a | 4 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 |
| 1b | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 2 | 4 |
| 2a | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 3 |
| 2b | 4 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 3a | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 3b | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 3 |
| 4a | 4 | 0 | 3 | 3 | 0 | 4 | 3 | 0 | 4 | 7 |
| 4b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 5a | 2 | 3 | 0 | 0 | 3 | 2 | 0 | 3 | 2 | 5 |
| 5b | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 4 |
| 6a | 5 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 1 | 5 |
| 6b | 0 | 4 | 4 | 4 | 2 | 2 | 2 | 2 | 2 | 8 |
| 7a | 5 | 4 | 0 | 0 | 5 | 4 | 0 | 5 | 4 | 9 |
| 7b | 6 | 0 | 0 | 0 | 4 | 2 | 4 | 2 | 0 | 6 |
| 8a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 8b | 0 | 9 | 0 | 0 | 9 | 0 | 0 | 6 | 0 | 9 |
| 9a | 8 | 0 | 0 | 4 | 4 | 0 | 4 | 4 | 0 | 8 |
| 9b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| 10a | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 10b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 11a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 11b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 12a | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 7 |
| 12b | 4 | 4 | 0 | 4 | 4 | 0 | 4 | 4 | 0 | 8 |
| Total | 31 | 30 | 15 | 15 | 32 | 24 | 17 | 27 | 19 | 100 |
| Paper cut score | 31 | 61 | 76 | 15 | 47 | 71 | 17 | 44 | 63 | |

Note: Examinees answer only half of the questions in Section B, i.e., from 7a to 12b. So only half of the points in Section B count.

## 4.3 Cut Scores Set by Two Panels Using the Same Performance Level Descriptors

For students who offer MSCE Mathematics, the assigning of their work to the performance categories depends on their combined score on the two subtests (known as Paper 1 and Paper 2) for the subject. By design, Paper 1 is constructed easier than Paper

2, although the papers are weighted the same: each carries 100 marks (points). The two subtests are administered on different days. Candidates answer all questions on Paper 1. Paper 2 has two sections. Section A has six compulsory questions. Section B has six questions also, but candidates are required to attempt only three. Thus, in determining the cut scores for Paper 2, the sum of cut scores for Section A of the paper is added to half of the sum of item cut scores for Section B, since candidates answer only half of the questions in this section. When the cut scores for the two papers have been decided, the final cut scores for the subject are derived by adding the corresponding cut scores of the two papers.

Table 4.4 Comparison of Cut Scores Set by Two Panels

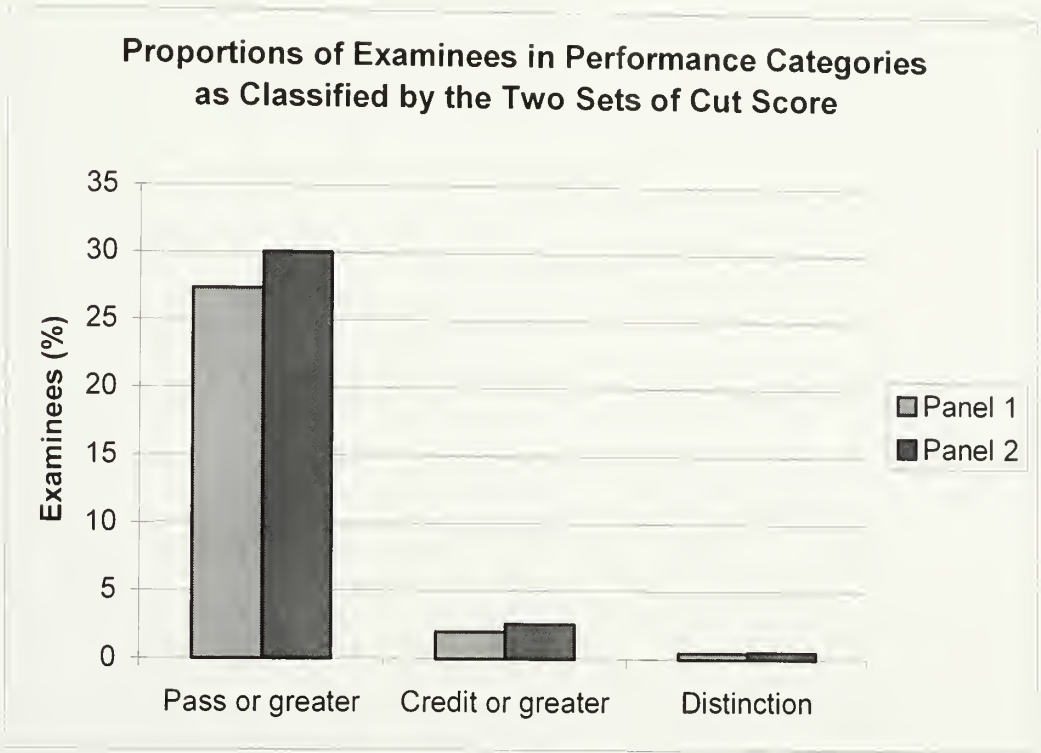
| | Pass | | Credit | | Distinction | |
|---|---|---|---|---|---|---|
| | Cut score | Examinees in this category and above (%) | Cut score | Examinees in this category and above (%) | Cut score | Examinees in the category |
| Panel 1 | 37 | 27.37 | 120 | 2.03 | 153 | 0.49 |
| Panel 2 | 34 | 30.07 | 114 | 2.57 | 151 | 0.55 |
| Average | 36 | 28.27 | 117 | 2.28 | 152 | 0.53 |
| Difference | 3 | 2.7 | 6 | 0.54 | 3 | 0.06 |
| Item rating Correlations | .796 | | .842 | | .905 | |

NB: The average cut scores have been rounded to the nearest whole number.

One of the objectives of this study was to determine the effect of using the same performance level descriptors on the cut scores set by two different panels. To investigate this, Panels 1 and 2 were requested to separately set cut scores on the two papers of the

2003 MSCE Mathematics. For each performance level, the cut scores for the two papers were combined, as is the practice during awards meetings. Table 4.4 shows the proportion of examinees assigned to each performance category using the cut scores set by the two panels. Item ratings by the two panels at each performance standard were correlated to determine level of agreement in their item ratings. Cut score differences were also computed. The proportions of examinees in the performance categories for the two years are also shown in Figure 4.1.

Figure 4.1




It is observed from Table 4.4 that all Panel 1 cut scores were a little higher than those set by Panel 2 but they appeared close, especially for two of the three cut scores. On a 200-point scale, a 3-point difference is only 1.5% on the test score scale.

Table 4.5 2003 Paper 1 Item Ratings by Panel 1 and Panel 2

| | Panel 1 | | | Panel 2 | | | Item Max. Score |
|---|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction | |
| 1 | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 2 | 3 | 0 | 0 | 3 | 0 | 0 | 3 |
| 3 | 0 | 0 | 3 | 0 | 0 | 3 | 3 |
| 4 | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 5 | 3 | 0 | 0 | 3 | 0 | 0 | 3 |
| 6 | 0 | 3 | 0 | 1 | 2 | 0 | 3 |
| 7 | 3 | 0 | 0 | 1 | 2 | 0 | 3 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0[1] | 4 |
| 9 | 0 | 4 | 0 | 0 | 2 | 2 | 4 |
| 10 | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 11 | 4 | 0 | 0 | 4 | 0 | 0 | 4 |
| 12 | 0 | 4 | 0 | 2 | 0 | 2 | 4 |
| 13 | 2 | 0 | 0 | 3 | 0 | 0 | 6 |
| 14 | 1 | 4 | 0 | 0 | 5 | 0 | 5 |
| 15 | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 17 | 2 | 0 | 0 | 0 | 0 | 0 | 5 |
| 18 | 2 | 0 | 2 | 2 | 0 | 0 | 5 |
| 19 | 0 | 5 | 0 | 0 | 5 | 0 | 5 |
| 20 | 0 | 2 | 4 | 2 | 0 | 2 | 6 |
| 21 | 0 | 0 | 0 | 0 | 5 | 0 | 5 |
| 22 | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| 23 | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| 24 | 2 | 0 | 3 | 0 | 0 | 5 | 5 |
| Total | 22 | 36 | 20 | 21 | 35 | 22 | 100 |
| Paper cut score | 22 | 58 | 78 | 21 | 56 | 78 | |

Table 4.6 2003 Paper 2 Item Ratings by Panel 1 and Panel 2

| | Panel 1 | | | Panel 2 | | | Item Max. Score |
|---|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction | |
| 1a | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 1b | 0 | 6 | 0 | 0 | 6 | 0 | 6 |
| 2a | 0 | 5 | 0 | 0 | 5 | 0 | 5 |
| 2b | 1 | 2 | 0 | 1 | 2 | 0 | 3 |
| 3a | 5 | 0 | 0 | 5 | 0 | 0 | 5 |
| 3b | 0 | 4 | 0 | 0 | 4 | 0 | 4 |
| 4a | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 4b | 0 | 3 | 0 | 0 | 3 | 0 | 3 |
| 5a | 3 | 0 | 0 | 2 | 0 | 0 | 7 |
| 5b | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| 6a | 2 | 1 | 0 | 3 | 0 | 0 | 3 |
| 6b | 2 | 0 | 4 | 1 | 0 | 5 | 6 |
| 7a | 2 | 2 | 0 | 2 | 2 | 0 | 4 |
| 7b | 0 | 0 | 2 | 0 | 3 | 4 | 11 |
| 8a | 0 | 9 | 0 | 0 | 9 | 0 | 9 |
| 8b | 0 | 6 | 0 | 0 | 6 | 0 | 6 |
| 9a | 0 | 6 | 0 | 0 | 6 | 0 | 6 |
| 9b | 0 | 0 | 9 | 0 | 2 | 7 | 9 |
| 10a | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 10b | 0 | 11 | 0 | 0 | 11 | 0 | 11 |
| 11a | 0 | 6 | 0 | 0 | 0 | 0 | 6 |
| 11b | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 12a | 0 | 5 | 0 | 0 | 5 | 0 | 5 |
| 12b | 2 | 0 | 0 | 0 | 0 | 0 | 10 |
| Total | 15 | 46.5 | 13.5 | 13 | 45 | 14.5 | 100 |
| Paper cut score | 15 | 61.5 | 75 | 13 | 58 | 72.5 | |

Note: The total is based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

Because Panel 1 cut scores were slightly higher than those of Panel 2, the proportions of examinees falling in each performance category were a little higher for Panel 2 than for Panel 1 as shown in Figure 4.1. However, the item rating correlations are significant ($p < .01$) at all performance standards. This provides evidence that there was a high degree of agreement between the two sets of item ratings at all the performance standards. Tables 4.5 and 4.6 provide detailed item ratings.

## 4.4 Consistency of Standards Over Years

To determine if application of the same performance level descriptors can yield comparable results on different forms of the test, the 2002 examination results were compared with those of 2003. As it has already been pointed out, the cut scores for the two forms of the examination were determined using the same criteria – the performance level descriptors. This use of the same criteria controlled the difference due to test difficulty.

The comparison involved computing proportions of examinees falling in each performance category. The score distributions for the two years were used for this purpose. (The 2003 score distribution was available at the time of data analysis but not earlier during the standard setting process.) An assumption was made that the two cohorts were of equal quality. To increase the likelihood that this assumption was not violated, ten well-established and stable schools were identified, and cut scores for the two years were used to determine the proportion of examinees that fell in each performance category. Thus, having controlled for test difficulty and assuming equivalence of students' ability, it was expected that the cut scores would produce approximately equal

proportions of students falling in each performance category. Table 4.7 presents the results for all examinees. The cut scores for the two years and the proportions of examinees in each category are also shown pictorially in Figure 4.2 and 4.3. Table 4.8 gives results for the ten schools, while Tables 4.9 and 4.10 give detailed 2002 item ratings.

Table 4.7 Comparison of Examination Results for All Examinees for 2002 and 2003

| Year | Pass and Above | | Credit and Above | | Distinction | |
|---|---|---|---|---|---|---|
| | Cut score | Examinees (%) | Cut score | Examinees (%) | Cut score | Examinees (%) |
| 2002 | 42 | 22.71 | 107 | 2.38 | 145 | 0.39 |
| 2003 | 36 | 28.27 | 117 | 2.28 | 152 | 0.53 |
| Difference | 6 | -5.56 | -10 | 0.10 | -7 | -.14 |

Figure 4.2

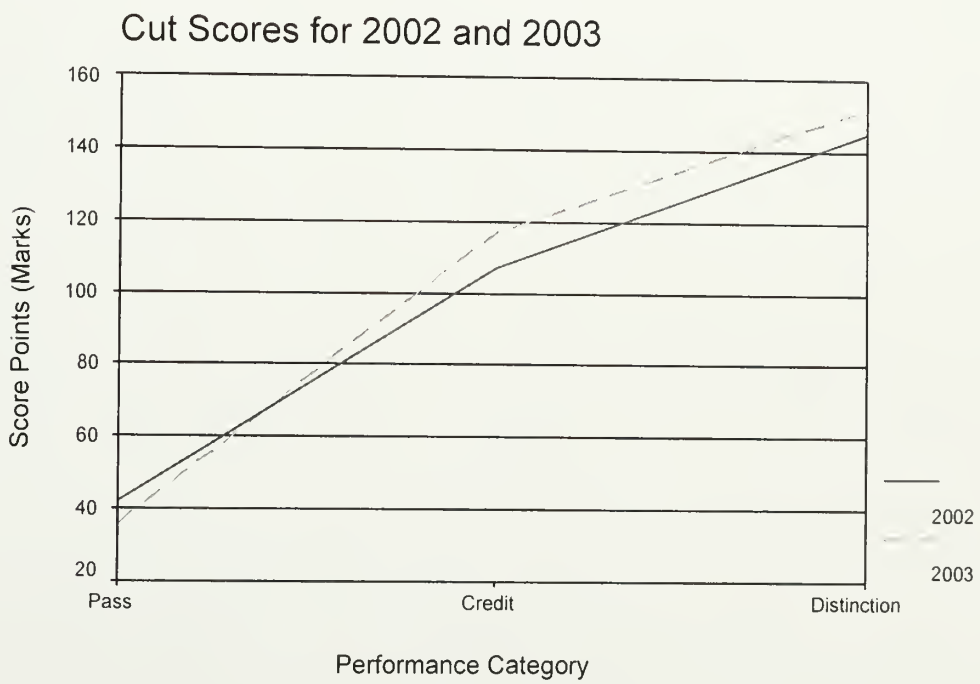
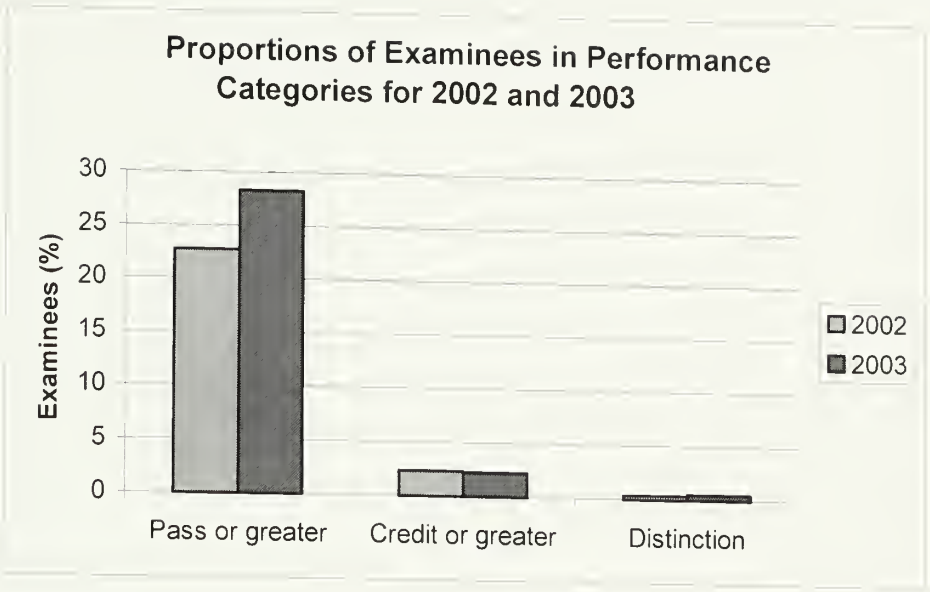


Figure 4.3




The results show that the cut scores for the two years were somewhat different, and varied across cut scores. The pass cut score for 2002 was greater than the 2003 cut score by six points. But for credit and distinction cut scores, those for 2003 were greater by ten and seven points, respectively. However, when the proportions of examinees falling in each performance category were considered, the credit and distinction cut scores produced almost similar results (see Figure 4.3), suggesting that the credit and distinction cut scores for the two years, though numerically different, might represent about the same level of proficiency, assuming equivalence of cohorts. Another possible explanation is that there are very few examinees in these performance regions so that any cut score would produce almost the same result. Thus, although the test forms for different years were assumed to be of equal difficulty, the fact that the panelists produced different cut scores when using the same criteria – the performance level descriptors – may have meant that they judged the test forms to be of different difficulty levels.

The observation that the proportions falling in each category were almost similar for credit and distinction categories but different for pass category needed further investigation. Perhaps part of the explanation is that so few candidates were in the higher scoring region. It may well have been that almost any cut scores would have generally produced the same percent of candidates in these categories. The difference in the proportions of candidates in the pass category may be due to the inclusion in the 2002 examination of some questions constructed from JCE work, but presented in a more complicated way than they would normally be presented for JCE candidates. The tendency was for the judges to classify such items under the pass category, believing that they were easy, having been taken from work of a lower educational level. There were no questions from JCE work in the 2003 MSCE Mathematics papers. It would therefore appear that the main reason for the pass cut scores for the two years to produce different proportions of examinees falling in the pass category was because the judges categorized some of the difficult questions from JCE syllabus under pass, because they were assumed to be easy. Judging from the item p-values (see Appendix F), they were not as easy as the judges had assumed. Had the judges considered the complexity of the presentation of such questions, they would have classified some of them under higher performance categories, and the cut score for pass would have been reduced, thus allowing some of those who failed to pass. The pass proportions would have become comparable over the two years. These results also suggested that the judges did not use the item p-values for the 2002 examination, although they had them. Apparently, they did not know what they meant, and so could not use them. Clearly, this was a problem and would need to be corrected in future standard setting studies.

Table 4.8

Comparison of Results for 2002 and 2003 for Ten Stable Schools

| | FAIL | | | | PASS AND ABOVE | | | | CREDIT AND ABOVE | | | | DISTINCTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| School | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 | 2002 | 2003 |
| | # students | # students | % | % | # students | # students | % | % | # students | # students | % | % | # students | # students | % | % |
| 1 | 74 | 50 | 71.2 | 54.3 | 30 | 42 | 28.9 | 45.7 | 8 | 6 | 7.7 | 6.5 | 1 | 1 | 1.0 | 1.1 |
| 2 | 44 | 34 | 32.6 | 24.6 | 104 | 97 | 75.4 | 70.3 | 24 | 40 | 17.8 | 29.0 | 5 | 16 | 3.7 | 11.6 |
| 3 | 38 | 64 | 13.3 | 24.0 | 248 | 203 | 86.7 | 76.0 | 91 | 89 | 31.8 | 33.3 | 35 | 53 | 12.2 | 19.9 |
| 4 | 86 | 68 | 70.5 | 48.9 | 36 | 71 | 29.5 | 51.1 | 4 | 12 | 3.3 | 8.6 | 1 | 1 | 0.8 | 0.7 |
| 5 | 62 | 30 | 73.8 | 46.1 | 22 | 35 | 26.2 | 53.9 | 4 | 4 | 4.8 | 6.2 | 1 | 1 | 1.2 | 1.5 |
| 6 | 39 | 29 | 48.2 | 36.7 | 42 | 50 | 51.9 | 63.3 | 10 | 13 | 12.4 | 16.5 | 0 | 5 | 0.0 | 6.3 |
| 7 | 153 | 121 | 52.9 | 56.5 | 136 | 93 | 47.1 | 43.5 | 14 | 26 | 4.8 | 12.1 | 3 | 8 | 1.0 | 3.7 |
| 8 | 108 | 57 | 75.0 | 53.3 | 36 | 50 | 25.0 | 46.7 | 4 | 2 | 2.8 | 1.9 | 2 | 0 | 1.4 | 0.0 |
| 9 | 98 | 110 | 63.2 | 71.4 | 57 | 44 | 36.8 | 28.6 | 8 | 3 | 5.2 | 2.0 | 3 | 0 | 1.9 | 0.0 |
| 10 | 86 | 53 | 68.8 | 53.5 | 39 | 46 | 31.2 | 46.5 | 13 | 7 | 10.4 | 7.1 | 3 | 0 | 2.4 | 0.0 |
| Total | 788 | 616 | | | 750 | 731 | | | 180 | 202 | | | 54 | 85 | | |
| % students | 51.2 | 45.7 | | | 48.8 | 54.3 | | | 11.7 | 15.0 | | | 3.5 | 6.3 | | |

Table 4.9 2002 Paper 1 Cut Scores

| Item # | Pass | Credit | Distinction | Item Maximum Score |
|---|---|---|---|---|
| 1 | 4 | 0 | 0 | 4 |
| 2 | 4 | 0 | 0 | 4 |
| 3 | 3 | 0 | 0 | 3 |
| 4 | 0 | 0 | 0 | 6 |
| 5 | 0 | 3 | 0 | 3 |
| 6 | 0 | 3 | 0 | 3 |
| 7 | 0 | 6 | 0 | 6 |
| 8 | 0 | 4 | 0 | 4 |
| 9 | 0 | 0 | 0 | 4 |
| 10 | 0 | 0 | 3 | 3 |
| 11 | 0 | 0 | 0 | 5 |
| 12 | 0 | 0 | 4 | 4 |
| 13 | 0 | 4 | 0 | 4 |
| 14 | 2 | 3 | 0 | 5 |
| 15 | 2 | 0 | 2 | 4 |
| 16 | 2 | 1 | 0 | 3 |
| 17 | 1 | 1 | 0 | 5 |
| 18 | 1 | 0 | 3 | 4 |
| 19 | 0 | 5 | 0 | 5 |
| 20 | 0 | 4 | 0 | 4 |
| 21 | 0 | 0 | 3 | 3 |
| 22 | 0 | 4 | 0 | 4 |
| 23 | 6 | 0 | 0 | 6 |
| 24 | 0 | 0 | 4 | 4 |
| Total | 25 | 38 | 19 | 100 |
| Paper cut score | 25 | 63 | 82 | |

Table 4.10 2002 Paper 2 Cut Scores

| Item # | Pass | Credit | Distinction | Item Maximum Score |
|---|---|---|---|---|
| 1a | 4 | 0 | 0 | 4 |
| 1b | 0 | 0 | 2 | 4 |
| 2a | 0 | 0 | 3 | 3 |
| 2b | 0 | 4 | 0 | 4 |
| 3a | 0 | 4 | 0 | 4 |
| 3b | 0 | 0 | 3 | 3 |
| 4a | 3 | 0 | 4 | 7 |
| 4b | 0 | 0 | 0 | 4 |
| 5a | 0 | 3 | 2 | 5 |
| 5b | 0 | 0 | 0 | 4 |
| 6a | 2 | 0 | 1 | 5 |
| 6b | 2 | 2 | 2 | 8 |
| 7a | 0 | 5 | 4 | 9 |
| 7b | 4 | 2 | 0 | 6 |
| 8a | 0 | 0 | 0 | 6 |
| 8b | 0 | 6 | 0 | 9 |
| 9a | 4 | 4 | 0 | 8 |
| 9b | 0 | 0 | 0 | 7 |
| 10a | 0 | 0 | 0 | 10 |
| 10b | 0 | 0 | 0 | 5 |
| 11a | 0 | 0 | 0 | 10 |
| 11b | 0 | 0 | 0 | 5 |
| 12a | 0 | 7 | 0 | 7 |
| 12b | 4 | 4 | 0 | 8 |
| Total | 17 | 27 | 19 | 100 |
| Paper cut score | 17 | 44 | 63 | |

Note: Note: The total is based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

The same comparison was made on the ten well-established and stable schools. The results are as shown in Table 4.8. The proportions of students in the various performance categories for the two years are more different for the ten schools than for all students. Thus, contrary to the assumption, there appears to be more instability in the ten schools than in the whole population. In fact, the instability is even greater for the individual schools. The possibility exists, therefore, that the stable schools were not really stable, at least for the two-year period under study. However, there is one observation that is consistent with the whole population: the 2003 results showed growth over the 2002 results. But four schools – schools 2, 3, 7, and 9 - performed in the opposite direction for the pass category. The effect of the performance of these schools on the ten-school sample was quite substantial. Therefore, it was decided to explain the results in terms of the whole population rather than the ten schools.

## 4.5. Comparison of Equated Cut Scores Derived from Common Judges and Common Items

The study was also interested in showing whether judgmental equating could further improve the quality of results produced by performance level descriptors. Two sets of equated cut scores, one based on common items and another based on common judges, were compared. There were ten common items, five from each paper, with a total of 53 points (marks). There were also three common judges. Using the ratings on common items and the ratings by common judges, the 2003 cut scores were equated to the scale of 2002. A linear equating method was used to compute the equating cut scores that were based on common items. The means and standard deviations of the ratings of

common items for the two years were used to equate the scores. Table 4.11 presents these values.

Table 4.11 Means and Standard Deviations of the Ten Common Items for the Two Rating Occasions

| | 2002 | | | | | | 2003 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paper 1 | | | Paper 2 | | | Paper 1 | | | Paper 2 | | |
| | Pass | Cred. | Dist. | Pass | Cred. | Dist. | Pass | Cred. | Dist. | Pass | Cred. | Dist. |
| Mean | 1.00 | 2.20 | 0.80 | 0.80 | 2.40 | 1.400 | 0.98 | 2.16 | 0.78 | 0.88 | 2.50 | 1.02 |
| SD | 1.732 | 2.683 | 1.789 | 1.789 | 3.362 | 1.949 | 1.299 | 1.455 | 1.067 | 0.896 | 2.637 | 0.841 |

Note: These figures are for the category points, not cumulative points.

Each of the 2003 average item ratings by Panels 1 and 2 was equated to the 2002 scale using these values. The equated item ratings and the final paper cut scores for Papers 1 and 2 are presented in Tables 4.12 and 4.13, respectively.

As there were no standard deviations for the common judges in the 2002 cut scores (because judges reached panel consensus) a mean equating approach was used to derive equated cut scores. There were three common judges in a panel of six. The equated paper cut scores of the corresponding performance categories for the two papers were summed to get the final cut score. Comparison of the derived equated cut scores for the two equating approaches is shown in Table 4.16.

Table 4.12 Paper 1 Equated Item Ratings Based on Common Items

| | 2003 Paper 1 mean item ratings for Panels 1 & 2 | | | Equated item ratings to 2002 scale | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1 | 0 | 3 | 0 | 0.23 | 2.59 | 0.30 |
| 2 | 3 | 0 | 0 | 2.48 | 0.97 | 0.30 |
| 3 | 0 | 0 | 3 | 0.23 | 0.97 | 2.09 |
| 4 | 0 | 3 | 0 | 0.23 | 2.59 | 0.30 |
| 5 | 3 | 0 | 0 | 2.48 | 0.97 | 0.30 |
| 6 | 0.5 | 2.5 | 0 | 0.61 | 2.32 | 0.30 |
| 7 | 2 | 1 | 0 | 1.73 | 1.51 | 0.30 |
| 8 | 0 | 0 | 0 | 0.23 | 0.97 | 0.30 |
| 9 | 0 | 3 | 1 | 0.23 | 2.59 | 0.90 |
| 10 | 0 | 4 | 0 | 0.23 | 3.14 | 0.30 |
| 11 | 4 | 0 | 0 | 3.23 | 0.97 | 0.30 |
| 12 | 1 | 2 | 1 | 0.98 | 2.05 | 0.90 |
| 13 | 2.5 | 0 | 0 | 2.10 | 0.97 | 0.30 |
| 14 | 0.5 | 4.5 | 0 | 0.61 | 3.41 | 0.30 |
| 15 | 0 | 4 | 0 | 0.23 | 3.14 | 0.30 |
| 16 | 0 | 0 | 0 | 0.23 | 0.97 | 0.30 |
| 17 | 1 | 0 | 0 | 0.98 | 0.97 | 0.30 |
| 18 | 2 | 0 | 1 | 1.73 | 0.97 | 0.90 |
| 19 | 0 | 5 | 0 | 0.23 | 3.68 | 0.30 |
| 20 | 1 | 1 | 3 | 0.98 | 1.51 | 2.09 |
| 21 | 0 | 2.5 | 0 | 0.23 | 2.32 | 0.30 |
| 22 | 0 | 0 | 4 | 0.23 | 0.97 | 2.69 |
| 23 | 0 | 0 | 4 | 0.23 | 0.97 | 2.69 |
| 24 | 1 | 0 | 4 | 0.98 | 0.97 | 2.69 |
| Total | 21.5 | 35.5 | 21 | 21.65 | 42.46 | 19.79 |
| Paper Cut Score | 22 | 57 | 78 | 22 | 64 | 84 |

Note: Paper cut scores have been rounded off to the nearest whole number.

Table 4.13 Paper 2 Equated Item Ratings Based on Common Items

| | 2003 Paper 2 Average Item Ratings for Panels 1 & 2 | | | Paper 2 Equated Item Ratings to 2002 Scale | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1a | 0 | 3 | 0 | 0.48 | 1.98 | 0.48 |
| 1b | 0.5 | 5.5 | 0 | 0.73 | 3.23 | 0.48 |
| 2a | 0 | 5 | 0 | 0.48 | 2.98 | 0.48 |
| 2b | 0.5 | 2.5 | 0 | 0.73 | 1.73 | 0.48 |
| 3a | 5 | 0 | 0 | 2.98 | 0.48 | 0.48 |
| 3b | 0 | 4 | 0 | 0.48 | 2.48 | 0.48 |
| 4a | 0 | 0 | 0 | 0.48 | 0.48 | 0.48 |
| 4b | 0 | 3 | 0 | 0.48 | 1.98 | 0.48 |
| 5a | 2.5 | 0 | 0 | 1.73 | 0.48 | 0.48 |
| 5b | 0 | 0 | 4 | 0.48 | 0.48 | 2.48 |
| 6a | 2.5 | 0.5 | 0 | 1.73 | 0.73 | 0.48 |
| 6b | 1.5 | 0 | 4.5 | 1.23 | 0.48 | 2.73 |
| 7a | 2 | 2 | 0 | 1.48 | 1.48 | 0.48 |
| 7b | 0 | 1.5 | 3 | 0.48 | 1.23 | 1.98 |
| 8a | 0 | 9 | 0 | 0.48 | 4.99 | 0.48 |
| 8b | 0 | 6 | 0 | 0.48 | 3.48 | 0.48 |
| 9a | 0 | 6 | 0 | 0.48 | 3.48 | 0.48 |
| 9b | 0 | 1 | 8 | 0.48 | 0.98 | 4.48 |
| 10a | 0 | 0 | 0 | 0.48 | 0.48 | 0.48 |
| 10b | 0 | 11 | 0 | 0.48 | 5.99 | 0.48 |
| 11a | 0 | 3 | 0 | 0.48 | 1.98 | 0.48 |
| 11b | 0 | 0 | 0 | 0.48 | 0.48 | 0.48 |
| 12a | 0 | 5 | 0 | 0.48 | 2.98 | 0.48 |
| 12b | 1 | 0 | 0 | 0.98 | 0.48 | 0.48 |
| Total | 14 | 45.75 | 14 | 15.64 | 31.53 | 15.64 |
| Paper Cut Score | 14 | 60 | 74 | 16 | 47 | 63 |

Note: Paper cut scores have been rounded off to the nearest whole number.
The total is based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

In Tables 4.14 and 4.15 are 2003 average item ratings by Panel 4 (with common judges) for Paper 1 and Paper 2, respectively.

Table 4.14 2003 Paper 1 Average Item Ratings by Panel with Common Judges

| Item # | Pass | Credit | Distinction | Item Maximum Score |
|---|---|---|---|---|
| 1 | 0.0 | 3.0 | 0.0 | 3 |
| 2 | 3.0 | 0.0 | 0.0 | 3 |
| 3 | 0.0 | 0.0 | 3.0 | 3 |
| 4 | 1.0 | 2.0 | 0.0 | 3 |
| 5 | 3.0 | 0.0 | 0.0 | 3 |
| 6 | 0.0 | 3.0 | 0.0 | 3 |
| 7 | 0.3 | 1.7 | 0.0 | 3 |
| 8 | 0.0 | 1.3 | 0.0 | 4 |
| 9 | 1.3 | 2.7 | 0.0 | 4 |
| 10 | 0.0 | 4.0 | 0.0 | 4 |
| 11 | 4.0 | 0.0 | 0.0 | 4 |
| 12 | 0.0 | 2.7 | 1.3 | 4 |
| 13 | 2.7 | 0.0 | 0.0 | 6 |
| 14 | 0.7 | 4.3 | 0.0 | 5 |
| 15 | 0.0 | 4.0 | 0.0 | 4 |
| 16 | 0.3 | 0.3 | 1.0 | 5 |
| 17 | 0.3 | 1.3 | 3.3 | 5 |
| 18 | 3.0 | 1.7 | 0.0 | 5 |
| 19 | 0.0 | 5.0 | 0.0 | 5 |
| 20 | 0.3 | 0.3 | 0.0 | 6 |
| 21 | 0.0 | 3.3 | 0.0 | 5 |
| 22 | 0.0 | 0.0 | 0.0 | 4 |
| 23 | 0.0 | 0.0 | 4.0 | 4 |
| 24 | 1.3 | 1.3 | 2.3 | 5 |
| Total | 21.3 | 42.0 | 15.0 | 100 |
| Paper Cut Scores | 21 | 63 | 78 | |

Note: Paper cut scores have been rounded off to the nearest whole number.

Table 4.15 2003 Paper 2 Average Item Ratings by Panel with Common Judges

| Item # | Pass | Credit | Distinction | Item Maximum Score |
|---|---|---|---|---|
| 1a | 0.0 | 3.0 | 0.0 | 3 |
| 1b | 0.7 | 3.3 | 2.0 | 6 |
| 2a | 0.0 | 5.0 | 0.0 | 5 |
| 2b | 2.0 | 1.0 | 0.0 | 3 |
| 3a | 5.0 | 0.0 | 0.0 | 5 |
| 3b | 0.0 | 4.0 | 0.0 | 4 |
| 4a | 0.0 | 0.0 | 0.0 | 6 |
| 4b | 0.0 | 3.0 | 0.0 | 3 |
| 5a | 0.7 | 1.0 | 0.7 | 7 |
| 5b | 0.0 | 0.0 | 3.7 | 4 |
| 6a | 1.0 | 2.0 | 0.0 | 3 |
| 6b | 0.0 | 0.0 | 0.0 | 6 |
| 7a | 0.3 | 2.7 | 1.0 | 4 |
| 7b | 2.7 | 0.7 | 2.3 | 11 |
| 8a | 0.0 | 6.0 | 3.0 | 9 |
| 8b | 0.0 | 6.0 | 0.0 | 6 |
| 9a | 0.0 | 6.0 | 0.0 | 6 |
| 9b | 0.0 | 0.7 | 8.3 | 9 |
| 10a | 4.0 | 0.0 | 0.0 | 4 |
| 10b | 0.0 | 11.0 | 0.0 | 11 |
| 11a | 0.7 | 3.3 | 0.0 | 6 |
| 11b | 0.0 | 1.7 | 0.0 | 9 |
| 12a | 0.3 | 4.7 | 0.0 | 5 |
| 12b | 0.0 | 1.7 | 0.0 | 10 |
| Total | 13.3 | 44.5 | 13.7 | 100 |
| Paper Cut Score | 13 | 58 | 72 | |

Note: Paper cut scores have been rounded off to the nearest whole number. Totals are based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

Table 4.16 Comparison of Equated Cut Scores Derived from Common Items and Common Judges

| | Pass | Credit | Distinction |
|---|---|---|---|
| 2003 cut scores (Using PLDs only) | 36 | 117 | 152 |
| Common Items (10 common items) | 38 | 111 | 147 |
| Common Judges (3 common judges) | 34 | 121 | 150 |
| Absolute Cut Score Difference | 4 | 10 | 3 |

It is observed from these results that, except for the credit cut scores, the equated cut scores that are based on common items are close to those that are based on common judges, considering that the score scale extends up to 200 points. The equated pass cut scores by both equating approaches are lower than the 2002 cut score of 42 (see Table 4.7), providing additional information to suspect that the 2002 judges did, indeed, underestimate the difficulty of the items they rated pass.

The degree of discrepancy between the two sets of equated cut scores may also indicate the extent to which the descriptors are functioning. Put in another way, if the descriptors are working, and both the common items and common judges equating are implemented with small errors, they should lead to same results. The absolute differences are 4, 10, and 3 for pass, credit, and distinction respectively. Except for the credit cut score, and considering the small sample size and a large score scale of 200 points, these differences are not substantial and provide encouragement that common judges equating could also be useful in the future.

Since the cut scores for both 2002 and 2003 examination forms were determined using the same criteria – the performance level descriptors – the two sets of cut scores

can legitimately be said to be equivalent in terms of the proficiency levels they represent. The equivalence of cut scores could further be strengthened by using more data to generate them. On this account, the equated cut scores are more legitimate, because more data, besides PLDs, were used. For this reason, the equated cut scores that are based on common items are more legitimate because ten common items were used as compared to only three common judges.

## 4.6 Comparison of Ratings Before and After Scoring Students' Answers

The impact of SMEs' participation in the scoring of students' answers on the cut scores was also studied. To investigate this, some SMEs who had set cut scores four weeks earlier, were requested to do the exercise again after participating in the scoring exercise. Their results were compared with those that were set before participating in the scoring exercise. The summary of the results are presented in Table 4.17, and the detailed item ratings for Papers 1 and 2 are presented in Tables 4.18 and 4.19, respectively.

Table 4.17 Summary of Cut Scores Set Before and After Scoring Students' Answers

| Session | Pass | Credit | Distinction |
|---|---|---|---|
| Before Scoring | 36 | 117 | 152 |
| After Scoring | 34 | 119 | 148 |
| Difference | 2 | -2 | 4 |
| Correlations | .773 | .924 | .769 |

Table 4.18 2003 Paper 1 Average Item Ratings Before and After Scoring

| | Before Scoring | | | After Scoring | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1 | 0 | 3 | 0 | 0.5 | 2.5 | 0 |
| 2 | 3 | 0 | 0 | 3 | 0 | 0 |
| 3 | 0 | 0 | 3 | 0 | 0 | 3 |
| 4 | 0 | 3 | 0 | 0.5 | 2.5 | 0 |
| 5 | 3 | 0 | 0 | 3 | 0 | 0 |
| 6 | 0.5 | 2.5 | 0 | 0 | 3 | 0 |
| 7 | 2 | 1 | 0 | 0.17 | 1.83 | 0.5 |
| 8 | 0 | 0 | 0 | 0 | 0.67 | 0.17 |
| 9 | 0 | 3 | 1 | 0.67 | 3.33 | 0 |
| 10 | 0 | 4 | 0 | 0 | 4 | 0 |
| 11 | 4 | 0 | 0 | 4 | 0 | 0 |
| 12 | 1 | 2 | 1 | 1.33 | 1.33 | 1.33 |
| 13 | 2.5 | 0 | 0 | 2 | 0.17 | 0.17 |
| 14 | 0.5 | 4.5 | 0 | 0.33 | 4.67 | 0 |
| 15 | 0 | 4 | 0 | 0.67 | 3.33 | 0 |
| 16 | 0 | 0 | 0 | 0.17 | 0.17 | 0.67 |
| 17 | 1 | 0 | 0 | 0.33 | 1.67 | 2.50 |
| 18 | 2 | 0 | 1 | 2.33 | 1 | 0 |
| 19 | 0 | 5 | 0 | 0 | 5 | 0 |
| 20 | 1 | 1 | 3 | 0.17 | 0.33 | 1.33 |
| 21 | 0 | 2.5 | 0 | 0 | 3.33 | 0.83 |
| 22 | 0 | 0 | 4 | 0 | 0 | 0.67 |
| 23 | 0 | 0 | 4 | 0 | 0.67 | 3.33 |
| 24 | 1 | 0 | 4 | 1.33 | 0.67 | 3 |
| Total | 21.5 | 35.5 | 21 | 20.5 | 40.17 | 17.50 |
| Cut Score | 22 | 57 | 78 | 21 | 61 | 78 |

Table 4.19 2003 Paper 2 Average Item Ratings Before and After Scoring

| | Before Scoring | | | After Scoring | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1a | 0 | 3 | 0 | 0.00 | 3.00 | 0.00 |
| 1b | 0.5 | 5.5 | 0 | 0.33 | 4.17 | 1.50 |
| 2a | 0 | 5 | 0 | 0.00 | 5.00 | 0.00 |
| 2b | 0.5 | 2.5 | 0 | 1.67 | 1.33 | 0.00 |
| 3a | 5 | 0 | 0 | 5.00 | 0.50 | 0.00 |
| 3b | 0 | 4 | 0 | 0.17 | 3.33 | 0.00 |
| 4a | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 4b | 0 | 3 | 0 | 0.00 | 2.50 | 0.50 |
| 5a | 2.5 | 0 | 0 | 0.67 | 0.50 | 0.33 |
| 5b | 0 | 0 | 4 | 0.17 | 0.00 | 3.67 |
| 6a | 2.5 | 0.5 | 0 | 0.67 | 2.33 | 0.00 |
| 6b | 1.5 | 0 | 4.5 | 0.00 | 0.00 | 0.00 |
| 7a | 2 | 2 | 0 | 1.50 | 2.33 | 0.50 |
| 7b | 0 | 1.5 | 3 | 2.00 | 0.67 | 1.83 |
| 8a | 0 | 9 | 0 | 0.33 | 7.17 | 1.50 |
| 8b | 0 | 6 | 0 | 0.00 | 6.00 | 0.00 |
| 9a | 0 | 6 | 0 | 0.00 | 6.00 | 0.00 |
| 9b | 0 | 1 | 8 | 0.00 | 0.33 | 8.67 |
| 10a | 0 | 0 | 0 | 4.00 | 0.00 | 0.00 |
| 10b | 0 | 11 | 0 | 0.00 | 10.67 | 0.33 |
| 11a | 0 | 3 | 0 | 1.00 | 3.33 | 0.00 |
| 11b | 0 | 0 | 0 | 0.00 | 0.83 | 0.00 |
| 12a | 0 | 5 | 0 | 0.17 | 4.83 | 0.00 |
| 12b | 1 | 0 | 0 | 0.00 | 1.50 | 0.67 |
| Total | 14 | 45.75 | 14 | 13.17 | 44.50 | 12.75 |
| Cut Score | 14 | 60 | 74 | 13 | 58 | 70 |

Note: Paper cut scores have been rounded off to the nearest whole number.
Total are based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

It is clear from these results that the magnitude of difference between the cut scores before and after scoring students' answers was small, suggesting that participation in scoring did not have an impact on the cut scores. The correlations of item ratings at all performance standards were significant ($p < .01$) and high (See Table 4.17). There is a remote possibility that some panelists remembered how they rated the items the first time. It is much more believable that judges were simply consistent in their judgments over one month interval, with participation in scoring sharing little or no effect. However, since these results were produced by trained judges, it is possible that this level of consistency was the effect of training.

## 4.7 Comparison of Cut Scores Set by Trained and Untrained SMEs

Another intent of the study was to compare the cut scores set by trained and untrained subject matter experts (SMEs). To investigate this, the cut scores set by six SMEs (Panel 4) were compared with those set by four untrained SMEs (Panel 5). The two panels used the same performance level descriptors that were developed and used during the standard setting workshop. The untrained SMEs were briefed about what the exercise entailed. Comparison of the two sets of cut scores was made by computing panel cut score differences and standard deviations. Tables 4.20 presents a summary of the results, and Tables 4.21 and 4.22 compares the average item ratings.

Table 4.20 Comparison of Cut Scores Set by Trained and Untrained Judges

| | Pass | Credit | Distinction |
|---|---|---|---|
| Original | 36 | 117 | 152 |
| Trained | 34 | 119 | 148 |
| | (3.14) | (5.27) | (6.00) |
| Untrained | 37 | 128 | 163 |
| | (12.52) | (17.92) | (13.81) |
| Correlations | .813 | .919 | .846 |

NB: The figures in brackets are standard deviations.

Figure 4.4 shows wider difference in higher performance standards.

Figure 4.4

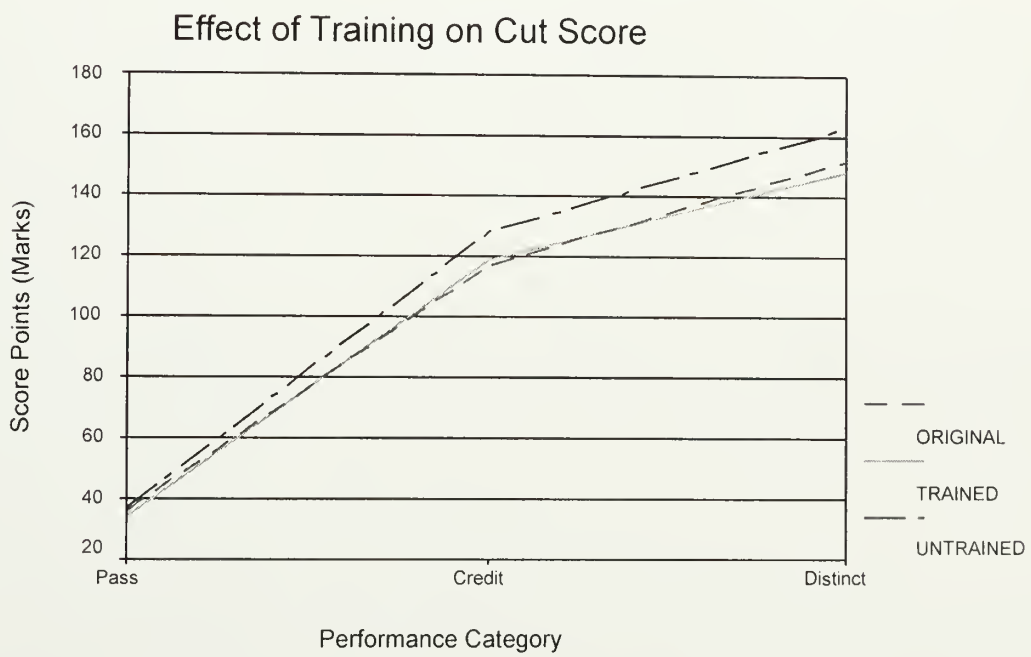


Table 4.21 Comparison of 2003 Paper 1 Average Item Ratings by Trained and Untrained SMEs

| | Trained SMEs | | | Untrained SMEs | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1 | 0.5 | 2.5 | 0 | 0.75 | 2.25 | 0 |
| 2 | 3 | 0 | 0 | 3 | 0.00 | 0 |
| 3 | 0 | 0 | 3 | 0.5 | 0.00 | 2.5 |
| 4 | 0.5 | 2.5 | 0 | 0 | 3.00 | 0 |
| 5 | 3 | 0 | 0 | 2 | 1.00 | 0 |
| 6 | 0 | 3 | 0 | 0.75 | 2.25 | 0 |
| 7 | 0.17 | 1.83 | 0.5 | 1 | 2.00 | 0 |
| 8 | 0 | 0.67 | 0.17 | 0 | 1.25 | 0.5 |
| 9 | 0.67 | 3.33 | 0 | 0 | 3.25 | 0.5 |
| 10 | 0 | 4 | 0 | 0.25 | 2.75 | 1 |
| 11 | 4 | 0 | 0 | 3.25 | 0.75 | 0 |
| 12 | 1.33 | 1.33 | 1.33 | 2.25 | 0.50 | 2.25 |
| 13 | 2 | 0.17 | 0.17 | 2 | 0.25 | 2.25 |
| 14 | 0.33 | 4.67 | 0 | 0.25 | 4.75 | 0 |
| 15 | 0.67 | 3.33 | 0 | 0 | 4.00 | 0 |
| 16 | 0.17 | 0.17 | 0.67 | 1.5 | 0.50 | 0.5 |
| 17 | 0.33 | 1.67 | 2.50 | 0.25 | 2.75 | 0.75 |
| 18 | 2.33 | 1 | 0 | 0.5 | 0.25 | 0.25 |
| 19 | 0 | 5 | 0 | 0.25 | 4.75 | 0 |
| 20 | 0.17 | 0.33 | 1.33 | 0.25 | 2.25 | 2 |
| 21 | 0 | 3.33 | 0.83 | 0.25 | 3.00 | 0.5 |
| 22 | 0 | 0 | 0.67 | 0 | 0.25 | 1 |
| 23 | 0 | 0.67 | 3.33 | 0.25 | 1.75 | 2 |
| 24 | 1.33 | 0.67 | 3 | 0 | 0.25 | 4.25 |
| Total | 20.5 | 40.17 | 17.50 | 19.25 | 43.75 | 20.25 |
| Cut Score | 21 | 61 | 78 | 19 | 63 | 83 |

Table 4.22 Comparison of 2003 Paper 2 Average Item Ratings by Trained and Untrained SMEs

| | Trained SMEs | | | Untrained SMEs | | |
|---|---|---|---|---|---|---|
| Item # | Pass | Credit | Distinction | Pass | Credit | Distinction |
| 1a | 0.00 | 3.00 | 0.00 | 0 | 3.00 | 0.00 |
| 1b | 0.33 | 4.17 | 1.50 | 1.5 | 4.50 | 0.00 |
| 2a | 0.00 | 5.00 | 0.00 | 0.25 | 4.75 | 0.00 |
| 2b | 1.67 | 1.33 | 0.00 | 0.75 | 2.25 | 0.00 |
| 3a | 5.00 | 0.50 | 0.00 | 4.5 | 0.50 | 0.00 |
| 3b | 0.17 | 3.33 | 0.00 | 0.75 | 3.00 | 0.25 |
| 4a | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.50 |
| 4b | 0.00 | 2.50 | 0.50 | 0 | 2.75 | 0.25 |
| 5a | 0.67 | 0.50 | 0.33 | 1.75 | 1.25 | 0.75 |
| 5b | 0.17 | 0.00 | 3.67 | 0.5 | 0.25 | 3.25 |
| 6a | 0.67 | 2.33 | 0.00 | 0.75 | 2.25 | 0.00 |
| 6b | 0.00 | 0.00 | 0.00 | 0.5 | 0.00 | 0.25 |
| 7a | 1.50 | 2.33 | 0.50 | 3.5 | 0.50 | 0.00 |
| 7b | 2.00 | 0.67 | 1.83 | 1.25 | 3.50 | 0.75 |
| 8a | 0.33 | 7.17 | 1.50 | 0.75 | 8.25 | 0.00 |
| 8b | 0.00 | 6.00 | 0.00 | 1 | 3.25 | 1.75 |
| 9a | 0.00 | 6.00 | 0.00 | 0 | 6.00 | 0.00 |
| 9b | 0.00 | 0.33 | 8.67 | 0.25 | 1.50 | 7.25 |
| 10a | 4.00 | 0.00 | 0.00 | 3 | 1.00 | 0.00 |
| 10b | 0.00 | 10.67 | 0.33 | 0.5 | 9.50 | 1.00 |
| 11a | 1.00 | 3.33 | 0.00 | 1.5 | 2.50 | 2.00 |
| 11b | 0.00 | 0.83 | 0.00 | 0.75 | 1.25 | 1.00 |
| 12a | 0.17 | 4.83 | 0.00 | 0 | 4.50 | 0.50 |
| 12b | 0.00 | 1.50 | 0.67 | 0.5 | 1.50 | 0.50 |
| Total | 13.17 | 44.50 | 12.75 | 17.75 | 46.13 | 15.50 |
| Cut Score | 13 | 58 | 70 | 18 | 65 | 80 |

Note: Paper cut scores have been rounded off to the nearest whole number.
The total is based on the sum of points from 1a to 6b and half of the points in Section B, i.e., from 7a to 12b.

It was observed from these results that the trained SMEs consistently set lower cut scores than the untrained. It was also observed that the trained SMEs produced cut scores that were closer to the original than were the cut scores set by the untrained SMEs. The original cut scores were set during the standard setting workshop. Another observation is that there was considerably greater inter-judge variability amongst the untrained than trained SMEs, judging by the values of standard deviations. However, the correlations reported in Table 4.20 were all significant ($p<.01$) and high, suggesting high level of agreement of the two panels' perceptions of the relative difficulty of the items, but the large standard deviations amongst the untrained judges indicated considerably more variability in their judgments. With more training and possibly more discussion among judges after initial ratings, it is likely that these differences could be reduced. Certainly at their current level, they are far too high to defend the cut scores that were set. These results clearly show the difference training of SMEs can make in the reliability of the cut scores. The results also suggest that, while the performance level descriptors play an important role in guiding the item rating process, they alone are not enough to produce defensible results: they need to be accompanied by actual training of judges to a standard setting study.

Another important general observation to make is that the items in Section B of Paper 2 are choice questions (questions 7 to 12). By virtue of being choice questions, they are supposed to be of approximately equal difficulty and should have been rated approximately the same by the judges. But, as the results have shown, these questions are not of equal difficulty. Clearly the topic of students' choice of questions deserves more

attention. Interestingly, choice is attractive to students but may also introduce an element of unfairness to them.

## 4.8 Results of the Evaluation Survey

As part of the evaluation process of the standard setting study, judges were asked to complete a modified version of an evaluation survey questionnaire prepared by Hambleton (2001) (see Appendix N). All 20 judges answered the questions, and the results are presented next.

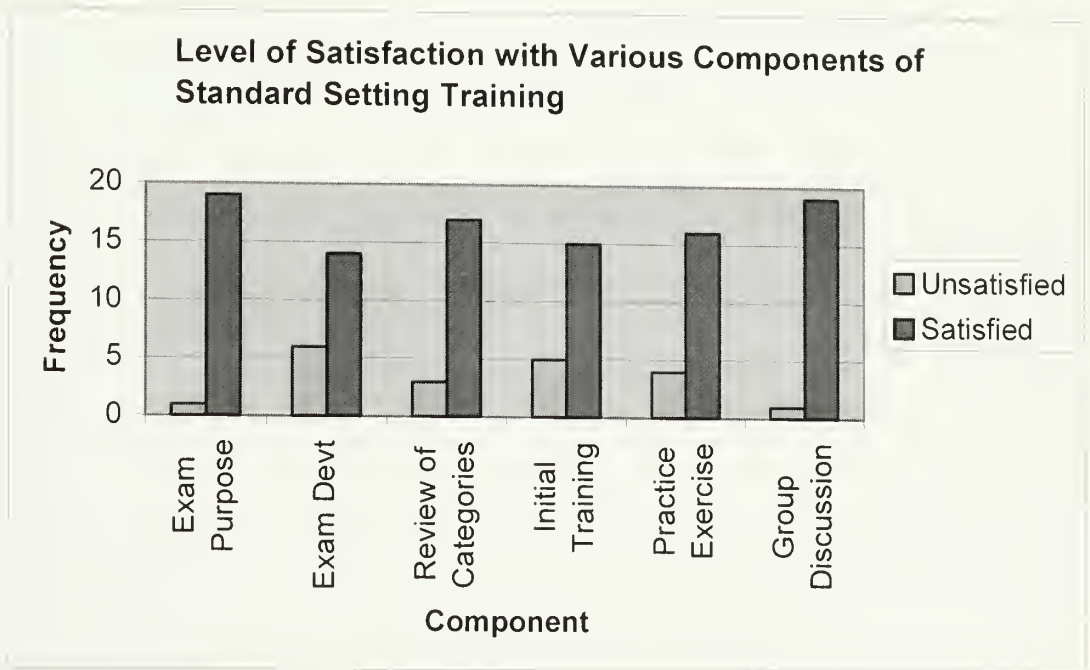
1. We would like your opinions concerning your level of satisfaction with the various components of the standard setting study. Place a tick (√) in the column that reflects your level of satisfaction of the various components of the standard setting study.

Table 4.23 Evaluation Results for Question 1

| Component | Not satisfied | Partially satisfied | Satisfied | Very satisfied |
|---|---|---|---|---|
| a. Description of the purpose of MSCE exam | 0 | 5% (1) | 45% (9) | 50% (10) |
| b. Description of the development of MSCE exam | 5% (1) | 25% (5) | 50% (10) | 20% (4) |
| c. Review of the four performance categories | 0 | 15% (3) | 55% (11) | 30% (6) |
| d. Initial training activities | 5% (1) | 20% (4) | 50% (10) | 25% (5) |
| e. Practice exercise | 0 | 20% (4) | 45% (9) | 35% (7) |
| f. Group discussion | 0 | 5% (1) | 20% (4) | 75% (15) |

The visual presentation in Figure 4.5 helps to see the level of satisfaction the judges had in the various components of the standard setting study. The responses in the first two columns, which represented lack of satisfaction, were combined and compared against combined responses of the last two columns, which represented general satisfaction with the component.

Figure 4.5




Note: Unsatisfied ^ Not Satisfied or Partially Satisfied.
Satisfied ^ Satisfied or Very Satisfied.

These results suggested that the judges were quite satisfied with the various components of the standard setting study. However, it was surprising to find that 30% of the judges were either not satisfied or only partially satisfied with the description of the development of MSCE examination. This is surprising in two ways: firstly because an opportunity was given for the judges to ask questions after the presentation so that

clarifications could be made; secondly because the development of MSCE examination is largely done by the teachers themselves. It is possible that this part of the presentation might have been rushed. Nevertheless, 70% of the judges were satisfied with the description of examination development.

2. Please rate the definitions provided during the training for these performance levels in terms of adequacy in guiding the standard setting process. Please CIRCLE one rating for each performance level.

Table 4.24 Evaluation Results for Question 2 (N = 19)

| Performance level | Adequacy of definitions | | | | |
|---|---|---|---|---|---|
| | Totally Inadequate (%) | | | | Totally adequate (%) |
| Fail | 0 | 10.5% (2) | 21.1% (4) | 31.6% (6) | 36.8% (7) |
| Pass | 0 | 10.5% (2) | 15.8% (3) | 47.4% (9) | 26.3% (5) |
| Credit | 0 | 5.3% (1) | 15.8% (3) | 52.6% (10) | 26.3% (5) |
| Distinction | 0 | 5.3% (1) | 5.3% (1) | 21.1% (4) | 68.4% (13) |

These results show that most of the judges found the definitions of the performance category adequate for standard setting.

3. How adequate was the training provided on the mathematics test booklet to prepare you to classify the students' performance?

Table 4.25 Evaluation Results for Question 3

| Rating | Percent |
|---|---|
| Totally adequate | 25% (5) |
| Adequate | 55% (11) |
| Somewhat adequate | 20% (4) |
| Totally inadequate | 0% (0) |

About 80% of the judges found the training on the test material adequate or totally adequate. The invitation letter asked the participants to familiarize themselves with the test material before going for the workshop. Furthermore, time was given during training for the participants to solve the problems on the test material. Clearly then, these findings are not surprising.

4. How would you judge the amount of time spent on training on the mathematics test booklet in preparing you to classify the students' performance?

Table 4.26 Evaluation Results for Question 4

| Rating | Percent |
|---|---|
| About right | 65% (13) |
| Too little time | 35% (7) |
| Too much time | 0% (0) |

No panelist thought the time for this activity was too much. Indeed, setting three cut scores on two papers required a substantial amount of time. Up to 35% of the panelists thought the time was too little. One reason for this finding was because this was

the first time the judges had carried out this type of activity. They probably needed more time to understand what to do and to actually carry out the exercise. The encouraging news is that the majority of the judges (65%) found the time adequate.

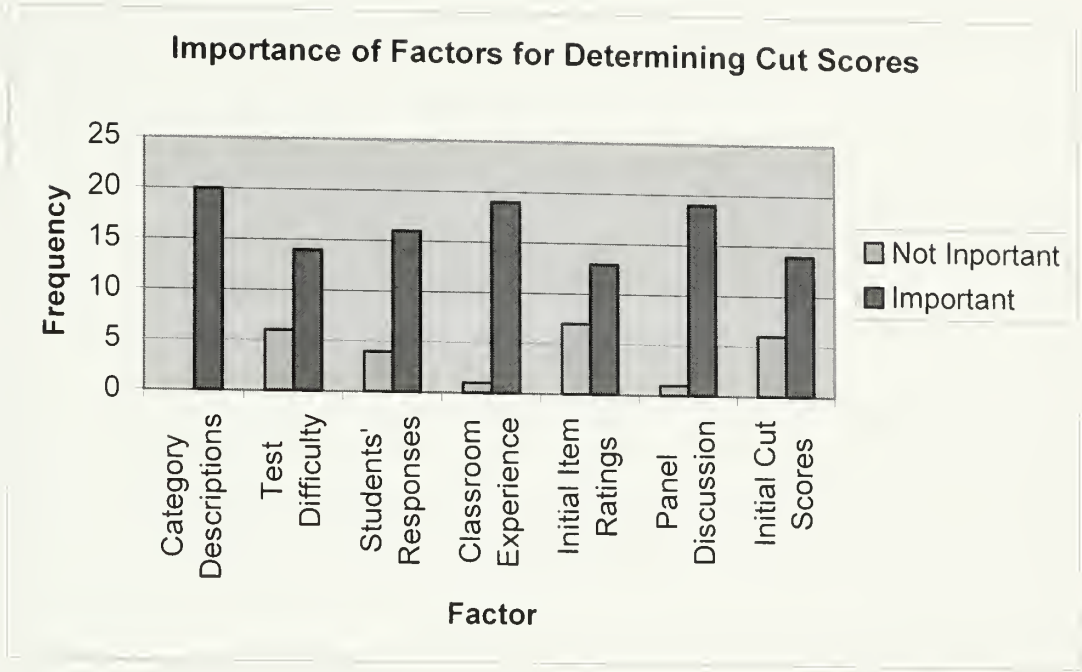
5. Indicate the importance of the following factors in determining the cut scores.

Table 4.27 Evaluation Results for Question 5

| Factor | Not important | Somewhat important | Important | Very important |
|---|---|---|---|---|
| a. The descriptions of Fail, Pass, Credit, Distinction | 0 | 0 | 45% (9) | 55% (11) |
| b. Your perceptions of the difficulty of the Mathematics Assessment material | 5% (1) | 25% (5) | 30% (6) | 40% (8) |
| c. Your perceptions of the quality of students' responses | 0 | 20% (4) | 40% (8) | 40% (8) |
| d. Your own classroom experience | 0 | 5% (1) | 35% (7) | 60% (12) |
| e. Your initial ratings of the items | 5% (1) | 30% (6) | 45% (9) | 20% (4) |
| f. Panel discussions | 5% (1) | 0 | 25% (5) | 70% (14) |
| g. The initial cut scores of other panelists | 10% (2) | 20% (4) | 35% (7) | 35% (7) |

Figure 4.6 gives a graphical display of the same information, with the first two columns (Not important and Somewhat important) collapsed into one category (Not important), and the last two columns (Important and Very important) also collapsed into one category (Important).

Figure 4.6




Note: Important ^ Not Important or Somewhat Important
Important ^ Important or Very Important

These results show that judges were generally satisfied with all the components of training. In particular, they considered the descriptions of performance categories, classroom experience, and panel discussions to be more important than the other components of determining the cut scores. It was surprising that a substantial number of judges (30%) did not consider their own perception of the difficulty of the test material important. This is surprising because the cut score position on the score scale depends very much on the perceived difficulty of the test material. It is also surprising that some judges did not consider the initial item ratings important, because subsequent changes to the item ratings also depended on initial ratings.

Table 4.28 Summary of Judges' Responses to the Evaluation Questions 6-13

| Question | Question | Percent (Frequency) |
|---|---|---|
| 6 | How would you judge the time allotted to do the first ratings of the questions on the test booklet? | |
| | About right | 70% (14) |
| | Too little time | 30% (6) |
| | Too much time | 0% (0) |
| 7 | How would you judge the time allotted to discuss the first set of panelists' ratings? | |
| | About right | 70% (14) |
| | Too little time | 30% (6) |
| | Too much time | 0% (0) |
| 8 | What confidence do you have in the classification of students at the DISTINCTION level? | |
| | Very high | 40% (8) |
| | High | 40% (8) |
| | Medium | 20% (4) |
| | Low | 0% (0) |
| 9 | What confidence do you have in the classification of students at the CREDIT level? | |
| | Very high | 30% (6) |
| | High | 45% (9) |
| | Medium | 25% (5) |
| | Low | 0% (0) |
| 10 | What confidence do you have in the classification of students at the PASS level? | |
| | Very high | 45% (9) |
| | High | 40% (8) |
| | Medium | 10% (2) |
| | Low | 5% (1) |

Table 4.28 Continued

| Question | Question | Percent (Frequency) |
|---|---|---|
| 11 | What confidence do you have in the classification of students at the FAIL level? | |
| | Very high | 45% (9) |
| | High | 40% (8) |
| | Medium | 10% (2) |
| | Low | 5% (1) |
| 12 | How confident are you that the Standard-Setting Method will produce a suitable set of standards for the performance levels: Pass, Credit, Distinction? | |
| | Very confident | 45% (9) |
| | Confident | 40% (8) |
| | Somewhat confident | 15% (3) |
| | Not confident at all | 0% (0) |
| 13 | How would you judge the suitability of the facilities for our study? | |
| | Highly suitable | 50% (10) |
| | Somewhat suitable | 35% (7) |
| | Not suitable at all | 15% (3) |

Although the majority of judges thought the time allotted to the various activities was sufficient, the fact that up to 30% of them said it was not, and no one said it was too much, suggests that more time would have been desirable. Regarding judges' confidence in the classification of students in the various performance categories, most judges indicated that they were either confident or very confident.

14. What strategy did you use to assign students to performance categories?

Through panel discussion, we decided where to place the cut score.

We were guided by the classification of the skills demanded by the questions.

We followed the processes involved in solving mathematical problems.

We considered capability of our own students.

We also changed some item cut scores after computing the final cut score for the examination

15. Were there any specific problems or exercises that were especially influential in your assignment of students to performance categories? If so, which ones?

No: (16)

Yes: Items from JC work (4)

16. Please provide us with your suggestions for ways to improve the standard-setting method and this workshop:

- It is an important exercise, and therefore needs a large group of participants with adequate time for discussion.
- Get views of other teachers and interested persons about the standards set to ensure they are accepted by all
- The exercise was somewhat rushed to save time and money. But more time was needed to do it thoroughly.
- The development of question papers should also consider the balance of the skills for various performance categories. If too many skills are concentrated in one performance category, this affects the proper derivation of the final cut score.

# CHAPTER 5

# DISCUSSION AND CONCLUSIONS

## 5.1 Introduction

The principle intent of this study has been to explore the viability of using performance level descriptors as common criteria for grading different cohorts of students or different tests. By using common grading criteria it was hoped that examination standards would be better maintained from year to year, and fairness across cohorts would be increased. This chapter describes some lessons learned from the study. The first section evaluates the whole process of standard setting as it was carried out. This is followed by discussion of the actual findings of the study. Conclusions, recommendations, and future research directions are offered at the end of the chapter.

## 5.2 Evaluation of the Standard Setting Process

The whole standard setting study will be evaluated in this section. The evaluation will be based on Kane's (1994) framework of procedural, internal and external evidence of validity.

### 5.2.1 Procedural Evidence

According to Plake (1998) cited by Carson (2001), one of the most successful ways of demonstrating the rationality and reasonableness of cut scores is evidence of procedural validity which focuses on who set the standards and how they did it. There are five types of procedural evidence – explicitness, practicability, implementation of

procedures, panelist feedback, and documentation. This study made an effort to satisfy all five types of evidence.

Explicitness. A clear step-by-step approach was followed as outlined in Chapter 3, and each step had to be understood, executed properly, and completed before moving on to the next step. The training of participants ensured that they understood the sequence of operations and what was expected of them. For example, the judges discussed the definitions of performance categories and practiced item rating before rating the examination items. Responses to the evaluation questionnaire provided corroborating evidence (see section 4.8). All these and other activities ensured that the criterion of explicitness was satisfied.

Practicability. For a standard setting method to satisfy the criterion of practicability, Berk (1986) asserted that the method must be easy to implement, the cut scores must be easy to compute, interpret, and be credible to lay people. The fact that the standard setting training and setting of cut scores took four days to do, and was carried out within the logistics of an academic exercise, demonstrate the practicability of the procedure. The judges did not have difficulties in computing the cut score. It involved the simple arithmetic operation of aggregating each judges' item ratings for each performance category and averaging across judges. Further, as it was described in the results chapter, the fact that they had to adjust the performance level descriptors following consideration of the consequence data shows that they were able to interpret the meaning of the cut score.

Implementation of procedures. Regarding the criterion of "implementation of procedures", Kane (2001) proposed the sub-criteria of selection of judges, training of

judges, definition of the performance standard, and quality of data collection. In this study, the judges were selected based on their expertise in the subject area of MSCE Mathematics and experience of teaching the subject and preparing students for the examination. The judges' teaching experience ranged from 3 to 29 years, with a mean of 8.3 years. The judges were selected from across the country, representing all major types of schools – public, grant-aided, and private. Some judges represented stakeholder institutions which included: the examiners, who are the Malawi National Examinations Board; the curriculum developers, who are the Malawi Institute of Education; and the end users such as the University of Malawi and Domasi College of Education. The majority of the judges (60%) were, at the time of the study, teaching the examination class. They had a lot of experience in the content area of MSCE Mathematics. Twenty percent of the judges were secondary school teacher trainers who also worked with the MSCE syllabus. In this regard, the sub-criterion of "selection of participants" was satisfied.

Another sub-criterion of "implementation of procedures" is training of judges. For good results, it is extremely important that the judges employed in the process are not only knowledgeable about the content, but also well trained in the method. They must fully understand the process they are to follow and what is required of them. Among other things, the judges need to be familiar with the measures that they will be working with, and understand the sequence of operations that they must perform. In this study, the training of judges included: description of purpose of the examination, how the examination is developed, processing of examination results, definitions of performance categories, misclassification errors, and practicing making item ratings. Every effort was made to meet all of the sub-criteria.

With respect to the definitions of the performance standards, participants discussed them at length. It was emphasized during training that the understanding of the definitions of performance standards was crucial to proper development of performance descriptors. This was the reason why the definitions were sent to judges two weeks ahead of the standard setting workshop. It was stressed during training that they should consider the definitions to represent the definition of a borderline performance, and their task would be to develop descriptors of borderline performance for each performance category. The results presented in Table 4.24 demonstrated the degree of adequacy of the definitions of the performance categories as perceived by the judges. Most of the judges found the definitions to be adequate for the standard setting study.

The sub-criterion of data collection is essentially about ensuring that appropriate and accurate data are collected. Ways of meeting this requirement include: balancing between absolute judgments and direct attention to passing rates (Shepard, 1980), reviewing decisions before finalizing the setting of cut scores, and consideration of the consequences of the cut score (Kane, 2001). As reported in the results chapter, cut scores were set three times, following discussion and consideration of the consequences of the cut score in terms of passing rates. The final cut scores were, indeed, a balance between absolute judgments and consideration of passing rates.

Judges' feedback. To satisfy the requirement of feedback, an evaluation questionnaire was distributed to the judges immediately after submitting their work. The proportions of frequencies of their responses were presented in the results chapter. The questions covered judges' level of satisfaction with the various components of the standard setting study (Table 4.23 and Figure 4.6), adequacy of time (Table 4.26),

definitions of the performance categories (Table 4.24), importance of certain factors in determining the cut score (Table 4.27 and Figure 4.7), and their suggestions to improve future standard setting studies.

Documentation. Since this study was conducted for academic purposes, and the results had to be reported in the form of a dissertation, the completion and production of this dissertation provide evidence that this requirement was satisfied.

### 5.2.2 Internal Evidence

Standard setting results that are not internally consistent do not justify any conclusions (Kane, 2001). It is therefore necessary to establish that the standard setting results are internally valid. Evidence for internal validity includes precision of estimates of the cut scores and consistency with empirical data such as item p-values. When the item p-values (see Appendix F) were compared with the item ratings in Tables 4.9 and 4.10, it was observed that in general, the items that were rated "pass" had higher p-values than those rated credit or distinction. There were some surprises, of course, but in the main, items classified as "pass" were considerably easier than those assigned to higher levels.

Internal evidence of validity can also be demonstrated by considering the cut scores for Paper 1 and Paper 2. It was mentioned that, by design, Paper 1 is easier than Paper 2. Therefore, Paper 1 should have higher cut scores than Paper 2. The results in Tables 4.2 and 4.3 confirm this. Paper 1 had the cut scores for pass, credit and distinction at 25, 63, and 82, respectively, while Paper 2 had its corresponding cut scores at 17, 44 and 63. The same picture is observed for the 2003 examination papers in Tables 4.5 and

4.6. In the 2003 data it was observed that both Panels 1 and 2 set credit cut scores a little higher for Paper 2 than Paper 1. The other cut scores were as expected.

Another way to address internal evidence is to compare the cut scores for randomly equivalent panels (Kane, 2001). In this study, two panels set cut scores on the 2003 examination. On a score scale of 200 points, the panel cut score differences were 3, 6, and 3 for pass, credit, and distinction, respectively. These differences were not substantial, considering the size of the score scale. The correlations of their item ratings were .796, .842, and .905 for pass, credit, and distinction, respectively. These were very high and significant ($p < .01$), suggesting high level of agreement.

### 5.2.3 External Evidence

One of the sources of external evidence of validity is reasonableness of the cut scores. It was learned from the results chapter that the performance level descriptors had to be adjusted twice to ensure that they produced reasonable and acceptable examination results to the judges. To achieve this, judges had to consider and discuss implications of impact data. The final cut scores were derived based on the judges' judgment and empirical data. Clearly then, there is some evidence, albeit limited, to address the requirement for external evidence of validity.

## 5.3 Discussion of Findings

The study set out to answer six research questions. The answers to the questions will be presented next.

### 5.3.1 Competences Necessary for Grading in a Particular Performance Category

One of the objectives of this study was to produce performance level descriptors which outlined the competences that examinees should demonstrate in order to be classified into particular performance categories. Used this way, performance level descriptors serve to guide and simplify judges' process of decision-making during cut score setting. If these guidelines are applied to different forms of the examination, then the resulting cut scores, which could be numerically different because of differences in the difficulty of the different forms of the examination, will represent roughly the same levels of proficiency in the subject area. This was the main intent of the study: to develop uniform criteria for assessing different cohorts of students to achieve justice, consistency, equity, and comparability in examination standards. The logic used was that consistency of examination standards would be achieved because cut scores for different forms of the examination will be derived from the same guidelines, which are the performance level descriptors. When consistency has been achieved, fairness will have been achieved as well. Of course, errors due to the process itself and the particular choices of judges still remain. These errors are inevitable.

Achieving consistency in examination standards has a very important advantage: it helps monitor growth or change in scores over time. There is a saying in the measurement field that: "if you want to measure change don't change the measure". Thus, to monitor educational growth, it is necessary that uniform grading criteria, i.e., same examination standards, be used over time. If standards are changed, it will be difficult to assess achievement growth.

Apart from helping achieve consistency, performance level descriptors can help improve classroom instruction. By specifying what students are expected to know and be able to do in order to achieve a particular grade category, performance level descriptors provide blueprints for what is important to teach and to learn. By making teachers and students know what is important to teach and to learn, performance level descriptors can have a powerful and positive effect on what goes on in the classroom. Using performance level descriptors will be like setting goals for students and teachers to reach. It is sensible to try to go where you want to go than to go where you do not know.

### 5.3.2 Comparison of Cut Scores Set by Two Panels Using the Same Performance Level Descriptors

One way to determine whether use of performance level descriptors can lead to consistent examination standards is to ask two or more panels to use them and see if they will come up with comparable cut scores. In this study, two panels set cut scores on the 2003 examination. Comparison of their cut scores involved computing differences (1) in their cut scores, (2) proportion of examinees classified in each performance category, and (2) correlations of their item ratings. The results were shown in Table 4.4. It was observed that cut scores set by Panel 1 were a little higher than those set by Panel 2. This represented a difference of 2.7% of candidates that would be classified differently if cut scores of one panel were replaced with those of the other panel. Compared to findings reported by Jaeger et al. (1980) and Good and Cresswell (1988) cited by Cresswell (1996), where up to 71% and 30% of examinees, respectively, could be classified differently, the cut scores set by the two panels in the current study were remarkably consistent. However, it should be noted that the panels in Jaeger and colleagues' study

represented different interest groups (teachers, school administrators, and counselors). The large differences in that study almost certainly reflected background differences of the panels. On the other hand, the closeness of panel results in the current study could be due to sharing of information during breaks. This would be unfortunate, but the possibility of this happening cannot be ruled out. Another reason could be the objective nature of the solution process in Mathematics. The quantitative nature of the discipline, and the objective solution process make the judgments less variable. In any case, the result is encouraging, but the generalizability of the result is unclear.

It is important to note also that the percentages of candidates affected by cut score differences will, besides difference in cut scores, depend on the score distribution. Small cut score differences at the center of a score distribution can result in larger proportions of students being affected than at the tail ends. In this current study for example, a difference of 3 marks (points) near the pass/fail boundary affected 2.7% of students, while a similar difference at the credit/distinction boundary affected less than 1% of the students. Recall too, that with the expected errors due to inconsistencies in ratings, some variation in the cut score would have been expected. That the errors are as small as they are, suggests that the performance level descriptors are quite helpful in reducing some of the judgmental error.

It seems, from this study, that performance level descriptors play a crucial role in ensuring some degree of consistency in item ratings. The high correlations and minor cut score differences are evidence for this. The results of this study, where two panels have produced almost similar cut scores, are consistent with the findings by Plake et al. (2000) and Kingston, et al. (2001), who found consistent results across panels.

### 5.3.3. Consistency of Standards Over Time

It was also the interest of this study to determine if the application of the same performance level descriptors on different forms of the examination would produce similar results. To investigate this, Panel 3 set cut scores on the 2002 MSCE Mathematics. The 2002 cut scores were compared with the average of those that were set by Panels 1 and 2 on the 2003 examination. The comparison involved computing proportions of examinees falling in each performance category, using examinees' score distributions for the two years. The results were presented in Table 4.7 in Chapter 4.

The results showed different cut scores at all the three thresholds. Since the same criteria were used to derive the cut scores for the two forms of the examination, this difference can only be explained by differences in examination difficulty as well as random error coming from the judges. The 2002 examination produced a pass cut score of 42 while the 2003 examination produced a pass cut score of 36. This means that the 2002 examination demanded more pass skills than the 2003 examination. One reason for this difference is because the 2002 examination contained some items from JCE syllabus. The 2003 examination did not have items from JCE syllabus. Judges made their own judgments as to which performance category they should classify such items, since there were no guidelines for classifying them. Judging from the way such items were classified, the tendency was for the judges to classify such items under pass, probably because they considered them to be easier. However, considering the complicated way some of the items were presented and their p-values, some of them were clearly under-rated and misclassified under pass. In other words, although the items came from JCE

topics, the actual mathematical maneuvering was more than JCE. The classification of such items under pass category raised the pass cut score for the 2002 examination.

It is important to explain why the 2002 examination had items from JCE work, while the 2003 examination did not. The 2002 MSCE examination was developed from the old examination syllabus, which assumed the JCE work. Inclusion of items from JCE work was, therefore, justified. The development of the 2003 examination, on the other hand, was guided by the new teaching syllabus, and inclusion of any questions outside the syllabus was unacceptable. It was unfortunate that the results were confounded by the JCE items but such are the problems that often arise in practice and that make experimental studies difficult to carry out.

When the credit and distinction cut scores were considered, the opposite picture was observed, where the 2002 examination had lower cut scores than the 2003 examination. But when the cumulative proportions of examinees falling in each performance category was considered, there was not much difference in the credit and distinction categories, despite wider differences in the cut scores. A possible explanation for this is that the credit and distinction cut scores for the two forms of the examination represented the same level of proficiency. This meant that performance level descriptors can be crucial in estimating equivalent cut scores on different forms of the examination test. In other words, although the cut scores for the two forms of the examination were different, they appeared to represent the same levels of proficiency. Another possible explanation is that there are so few examinees in the credit and distinction categories that any cut score would produce almost similar proportions of examinees in those categories.

The finding that judges set different cut scores on different forms of the examination has an important implication for test development. While an effort should always be made to develop parallel forms of the examinations, this finding has confirmed a well-documented psychometric assertion that it is extremely difficult to develop tests of exactly the same level of difficulty (Angoff, 1971; Newton, 1997; Norcini, 1990; Norconi & Shea, 1997). This means that the same level of proficiency can be represented by different cut scores on different forms of the test, or the same cut score can represent different levels of proficiency in different forms of the examination. However, if a test blueprint is used, and different forms of the examination are constructed based on the same blueprint, the problem of widely differing cut scores may be minimized.

Regarding the performance of the ten "stable schools" (Table 4.8), two observations can be made. The first one is that the stable schools performed much better than all students together, judging by the proportions of examinees in the performance categories. They were identified as well-established and stable schools on the basis that they are old and have been good schools. So, their quality of performance was to be expected. The second observation is that, although the general picture is better performance in 2003 than in 2002, the performance data in four schools was opposite. In addition, the drastic drop in the numbers of students that wrote the examination in these schools (1538 in 2002, 1347 in 2003), representing a drop of 12.4%, raise a question about the stability of these schools during the two years. For these reasons, the idea of stable schools was abandoned, and the performance of the whole population was used instead. Of course the presence of these and other schools that have performed in the

opposite direction to the general trend confounds the results, and complicates the explanation of the findings.

### 5.3.4 Comparison of Equated Cut Scores from Common Judges and Common Items

The study also compared equated cut scores that were based on common judges with those that were based on common items. The 2003 cut scores were equated to the 2002 scale using the two approaches. A linear equating method was used to compute the equating cut scores that were based on common items. But, as there were no standard deviations for the common judges in the 2002 cut scores (because judges reached panel consensus), a mean equating approach was used to derive equated cut scores. The results were presented in Table 4.16. As it was shown in the results chapter, the two approaches produced practically similar pass and distinction cut scores. Only the credit cut score was different. There are at least two reasons why the two equating approaches produced slightly different results. Firstly, the two equating approaches employed different statistical methods: linear equating for common items approach and mean equating for common judges approach. These approaches are not expected to produce similar results. Secondly, the common judges approach used only three common judges, which is really too small a number to produce a stable equating. Based on the present results, the common items equating would be recommended. However, the fact that even with few common judges, the two equating approaches have produced almost similar results, imply that with more common judges, better results would be expected.

A very important lesson that has been learned from this part of the study is that there is a rational procedure for estimating cut scores on different forms of an

examination. This has direct relevance to the MSCE examination whose certificates are valued the same, regardless of the year they were issued. By adjusting test scores based on how judges rate the same test items when they are on different forms of the examination, it is possible to obtain equivalent test scores on the two score scales. This is a very important exercise if examination standards are to be maintained. Since examinations will rarely be matched precisely in difficulty, and since equating scores is not currently being done, that leaves the need to equate examinations via the standard setting process.

### 5.3.5 Comparison of Ratings Before and After Scoring Students' Answers

The study was also interested to find out if participating in scoring students' answers affects standard setting judgments. To investigate this, judges who had set cut scores four weeks earlier, were requested to do the exercise again after participating in the scoring of candidates' answers. Their results were presented in Table 4.17. There were small cut score differences (cut score differences ranged from –2 to 4) in all the three cut scores, and the correlations were all significant ($p>.01$) and high. It can, therefore, be concluded from this study at least that participation in scoring students' work does not influence where the cut scores are set. Of course, this finding should be checked, and it certainly must be recognized that some training of scoring would be desirable during the standard setting training. One possible explanation for this finding is that the cut scores were set by trained judges who might have remembered how they rated the items the first time. Another explanation is that trained judges produce

consistent results over occasions (Raymond & Reid, 2001). So the consistency of cut scores might be due to training the judges received.

### 5.3.6 Comparison of Standards Set by Trained and Untrained SMEs

The cut scores produced by trained and untrained panelists were presented in Table 4.20. It was clear from the results that training of judges plays a very important role in determining the cut score. The untrained judges consistently produced higher cut scores than the trained panelists. Further, and significantly, the standard deviations of the cut scores produced by untrained judges were much higher than those produced by the trained panelists at all performance standards. However, the correlations between the ratings carried out by trained and untrained judges were significant at all performance levels. This means that there was less agreement among the untrained than the trained judges in the absolute difficulty level they assigned to the test items (using standard deviations), but they agreed strongly in the relative difficulty of test items (using correlations), a finding that is consistent with Lorge and Kruglov (1952, 1953) cited by Thorndike (1982). This also confirms the point that training of judges makes them have a common understanding of what it means to achieve a particular grade. It is a serious shortcoming if no or minimal standard setting training is given. It is highly unlikely that untrained judges will understand or correctly perform the required tasks without undergoing training.

## 5.4 Conclusions

MANEB administers MSCE and other examinations every year, and in each year, a different form of the examination is used. Consequently, cut scores are set on each form to decide who has reached the requirements for a particular performance category. Because similar grades from different forms are treated the same, the cut scores for each grade need to represent the same proficiency in the subject. This study has demonstrated that one way to increase comparability of grades from different forms is to use uniform criteria for grading the students. In this study, uniform criteria used were the performance level descriptors. The study has shown that performance level descriptors have the potential to reduce cut score variability due to ambiguity about the level of proficiency they represent. By using the same grading criteria, standards are more likely to be maintained and more equity between different cohorts of students will be achieved.

Conducting equating procedures can further enhance the role of performance level descriptors in ensuring equity and maintenance of standards. It is only when different forms of an examination have been equated that it becomes fair to treat similar grades on different forms of the examination the same. By equating grade cut scores for different forms of the examination, the grades awarded carry the same meaning in terms of the level of proficiency they represent. Consequently, equating helps to ensure consistency in examination standards. Equating is the preferred way to find equivalent scores on examinations over time because the stability of the equating is high, and all efforts can be planned into setting cut scores once on the baseline examination. Because test score equating was not possible due to disclosure of MSCE examination items after each administration, judgmental equating was used in this study, and has worked well for

common items approach, but not so well for common judges' approach. This was probably because only three common judges were used, as compared to ten common items.

Regarding the role of performance level descriptors in guiding the judges' judgmental process, the results consistently indicated a high degree of agreement amongst judges' judgments. If the content of the different forms of the examination is similar, the level of agreement is even higher, and the resulting cut scores on the different forms can safely be said to represent the same level of proficiency in the subject. However, due to the judgmental nature of the exercise, it is unlikely that any two panels will agree exactly. This study was no exception: minor differences were observed, but overall, the level of agreement was substantially high.

One of the requirements for a successful standard setting study is excellent training of the participants. Many authors have acknowledged the importance of this requirement (Cizek, 1996; Berk, 1996; Kane, 1998; Mills, 1995; Fehrmann et al., 1991, cited by Cresswell, 1996). It is further required that the quality of training itself be documented (Cizek, 1996, Hambleton, 2001). Thus, as part of the evaluation of the training, judges were asked to rate the quality of the various components of the training that was offered. The ratings were generally high. The findings of this study have also confirmed Raymond and Reid's (2001) contention that training improves stability of standards over occasions. The cut scores set by trained judges were almost identical to their original cut scores.

With respect to the hypothesized impact of scoring students' answers on setting cut scores, no impact was found. There was virtually no difference between the cut scores

set by judges who participated in the scoring activities and those who did not. A possible explanation for this is that data for this part of the study were collected from trained judges, and trained judges produce consistent results over occasion (Raymond & Reid, 2001). It should, all the same, be noted that there was no harm or benefit in participating in scoring before taking part in the standard setting activities. This finding opens up possibilities, for example, of using some scorers as standard setters.

The finding that the 2002 and 2003 forms of MSCE Mathematics had substantially different cut scores has implications for the test development process. The different cut scores were expected because the two forms cannot be constructed to be exactly equal in difficulty. But the observed differences were greater than had been expected. Probably different test specifications were used. But comparability of scores on different forms of assessments can only be justified if the assessments are similar in their tasks, cognitive demands, and conditions of administration. Use of test specifications is central to satisfying similarity in test tasks and cognitive demands.

### 5.4.1 Recommendations and Future Research Directions

The following recommendations are made based on the findings of this study.

Recommendation 1

The role that performance level descriptors play in standard setting and in maintenance of examination standards cannot be overemphasized. It is therefore recommended that MANEB embark on a campaign to produce performance level descriptors for all subject areas and for all examinations that it administers. More research is needed, however, on the best approach for preparing these descriptors.

Recommendation 2

The finding that trained panelists produce stable cut scores over occasion is good news for the maintenance of standards. It is therefore recommended that people who participate in standard setting or awards meetings, undergo standard setting training. Again, this study provides a starting point for just how this training might be provided.

Recommendation 3

For security reasons, MANEB does not reuse test items. This means that it is not possible to conduct test score equating. Because people treat similar grades from different forms of the examination the same, it is required to ensure that similar grades have the same meaning. The only viable option available is judgmental equating of examinations (see Hambleton, 2000). This study and the literature review have shown that judges are able to make judgments of item difficulty with sufficient accuracy. Their judgments can be used as the means for adjusting the difficulty of new forms of the examination to make them comparable to earlier forms. It is therefore recommended that MANEB conduct judgmental equating to ensure consistency of standards and comparability of grades

Recommendation 4

It has been shown in this study that different forms of MSCE examinations are not comparable, especially in terms of balancing the level of skills measured. It is therefore recommended that a test blueprint that will specify the weighting of skills of each performance category, guide the development and moderation of examination papers. "It is easier to try to be careful in constructing tests than it is to try to compensate for poor test construction afterward" (Potthoff, 1982, p. 202).

Recommendation 5

It was evident from the judges' item ratings that they did not take into account the item p-values that they were given. It is most likely that this was due to lack of psychometric knowledge. They did not understand the "p-values". It is therefore recommended that participants for future standard setting workshops be trained in the basics of psychometrics to understand item analysis results and specifically, item p-values.

Recommendation 6

As this study was conducted for academic purposes, and being the first of its kind, there were time and budget constraints. The evaluation results also alluded to the same shortcoming. The panel sizes were only barely adequate. It is recommended that further research with adequate resources and larger panels, with judges representing all important stakeholders, be conducted to improve the accuracy of the process of grading students.

Recommendation 7

The choice questions in Section B of paper 2 are assumed to be of equal difficulty. This is why each one of them carries 15 marks (points). However, the different item p-values and ratings by the judges appear to contradict this equal difficulty assumption. This finding may suggest that even when candidates sit for the same examination in general, candidates, based upon their selection of questions, actually take examinations of different difficulty. Furthermore, the candidates make their choice based on their perception of the difficulty of the questions. The perception is usually based on very hurriedly reading of the questions, which may reveal unforeseen difficulties when the answer is actually composed (Mathews, 1985). The decision is made during

examination with all its stress and pressure of time. It is therefore recommended that the policy of allowing candidates choice of questions be reviewed. It is further recommended that the review include a qualitative comparability of examination questions in order to determine the equivalence of choice questions, and when necessary, revisions to the scoring rubrics can be made to try and judgmentally equate the choice questions.

Recommendation 8

Examination results are frequently used as indicators of levels of students' achievement. But, as it has been shown, examination results can fluctuate widely. It is therefore recommended that a study be conducted to determine whether changes in examination pass rates correspond to actual changes in students' achievement levels. Do improved examination results mean higher achievement?

Recommendation 9

The standard setting method used in this study uses a pass-fail procedure for each score point. Inherent in this procedure is the assumption that the skills identified for borderline examinees are of the same difficulty level. Consequently, the method demanded 100% probability for borderline examinees to get items demanding borderline skills correct. Clearly, this assumption is not correct, and it is not expected that borderline examinees will find all borderline skills equally easy or difficult to perform. So, this method of determining cut score should not be viewed as final. Further improvements to the process of deriving cut scores are possible. It is therefore recommended that another study be conducted that will require the judges to use probability estimates of borderline performance on the skills in order to determine the item ratings.

Recommendation 10

A pass cut score of 13 on a 100-point test seems very low, as was observed in Paper 2. This was probably because the test did not include many pass skills. It is possible to have a respectable pass score by increasing the number of items assessing at that level. It is therefore recommended that a study be conducted to determine appropriate proportions of items assessing skills at each of the performance levels of the examination.

### 5.4.2 Final Remarks

The execution of this study was not without limitations. To do good research on standard setting requires adequate financial resources and time for the participants to deliberate and refine their proposed cut scores. Due to financial constraints, small sample sizes of participants were used. For the same reason, the standard setting participants could not be kept long enough to thoroughly refine their work. For these reasons, the findings of the study need to be treated with caution. Because of small sample sizes, the findings may reflect some instability. Therefore, further research is necessary to replicate the study with larger sample sizes and adequate time.

In spite of the constraints, the study has registered one important outcome: a program of standard setting has been started. The situation can only be hoped to get better. Setting a justifiable and valid cut score requires a rational process for determining it. If more people, representing different stakeholders such as subject matter experts, educators, parents, and policymakers, participate in the study, there will be greater acceptability of the outcome.

Finally, it is the desire of most examination authorities to use cut scores that separate those who have achieved from those who have not, with sufficient precision. The judges' decision accuracy can be improved by using some predetermined criteria for establishing cut scores. Further, use of criteria simplifies the judges' thinking process in determining the cut score, and can greatly help maintain examination standards if they are used consistently from year to year.

APPENDICES

APPENDIX A

2002 MSCE MATHEMATICS PAPER 1 (REPRODUCED WITH PERMISSION)

# THE MALAWI NATIONAL EXAMINATIONS BOARD

2002 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

MATHEMATICS

Subject Number: M131/I

**Tuesday, 15 October**

**Time allowed: 2 hours**

**9:00 - 11.:00 am**

## PAPER I

(100 marks)

**Instructions**

1. **This paper contains 7 pages. Please check.**
2. Answer **all** the **24** questions in this paper.
3. The maximum number of marks for each answer is indicated against each question.
4. Mathematical tables and answer books are provided.
5. Used supplementary sheets must be handed in together with the answer book.
6. All working must be clearly shown; it should be done on the same sheet as the rest of the answer.
7. Use of electronic calculators is not allowed.
8. Write your **Examination Number** on top of each page of your Answer Book.

**© 2002 MANEB**

**Turn Over**

Answer **all** the **twenty four** questions.

1. Simplify $\frac{(3\frac{1}{2} x 1\frac{1}{2}) - 3}{9}$ **(4 marks)**

2. **Figure 1** is a pie-chart representing sales of three commodities; tobacco, tea and coffee.

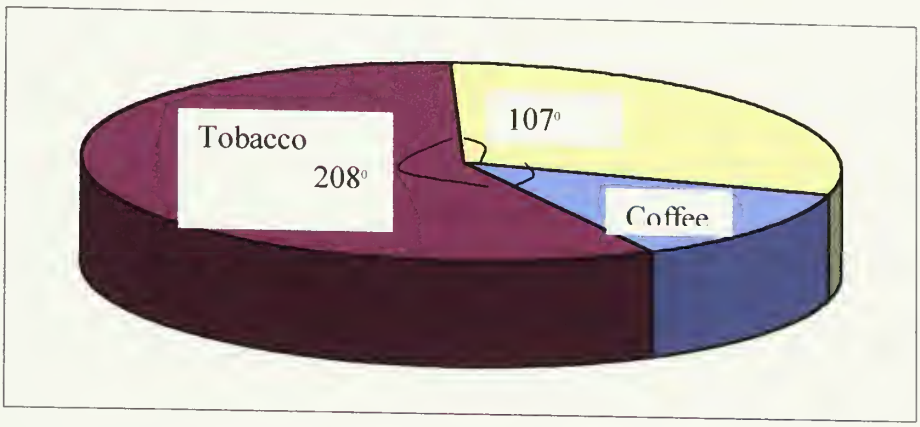



**Figure 1**

Express coffee sales as a percentage of the total sales **(4 marks)**

3. P is a point on the graph whose equation is $y = x^2 - 6x$. If the x-coordinate of P is 2, calculate its y coordinate. **(3 marks)**

4. In a cyclic quadrilateral ABCD twice angle BAD = three times angle DCB. Calculate angle BAD. **(6 marks)**

5. Given that angle $\theta$ is acute and that log cos $\theta$ = $\bar{1}.75$, evaluate $(\text{Cos}\,\theta)^2$. **(3 marks)**

6. Factorise completely, $1 - 16(1 - y)^2$ giving your answer in its simplest form. **(3 marks)**

**Continued/...**

7. In **Figure 2**, **DB** is perpendicular to the line **ABC**, **AE** = 25 cm, **BC** = 15 cm, angle **EAB** = 30°, and angle **BCD** = 45°.

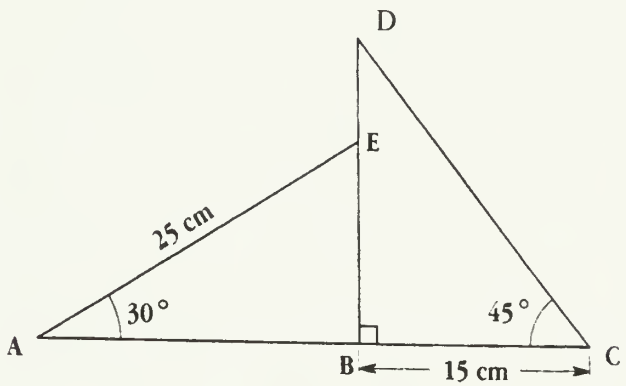



**Figure 2**

Calculate the length of **DE**. **(6 marks)**

8. Given that $\frac{a^7}{a^{-3}.a^2} = a^y$ **(4 marks)**

9. In **Figure 3**, **D** is the midpoint of the minor arc BDC, angle ABC = 40° and angle ACB = 60°.

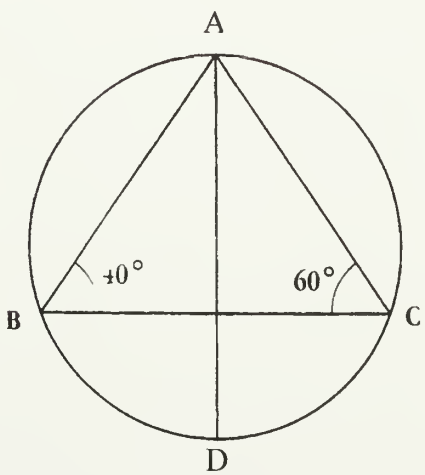



**Figure 3**

Calculate angle **DAC**. **(4 marks)**

**Continued →**

10. Make $x$ the subject of the formula $4y = a^x$. **(3 marks)**

11. In **Figure 4**, **O** is the centre of the centre of the circle, **TA** is a tangent, **BC** is parallel to **TA** and angle **BCT** = 37°.

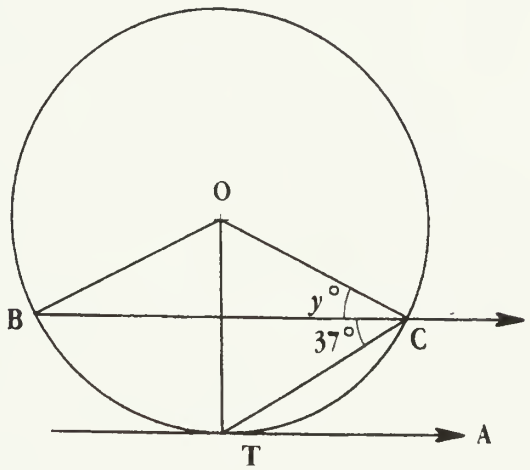



**Figure 4**

Calculate the value of the angle marked y. **(5 marks)**

12 Solve for $x$ $\log_x 125^{-1} = -3$ **(4 marks)**

13. Given that $2x$, $x$, $x + 3$, .....are terms in an Arithmetic Progression. Calculate the value of $x$.

**Continued**

13. In **Figure 5, AB** is parallel to **CD, EF and GH**. The parallel lines **AB, CD, EF** and **GH** intersect **QR** such that QX = XY = YR. **SU = 9** cm. **DU** = 8 cm and **TV** = 5 cm.

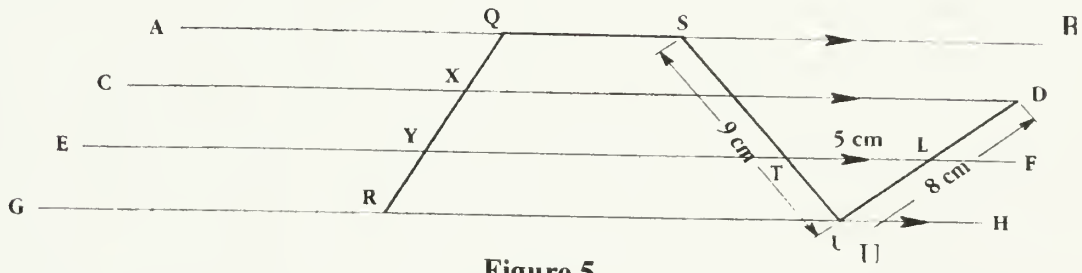



**Figure 5**

Prove that angle TUV = 90° **(5 marks)**

15. Simplify $\sqrt[3]{8a} + 2\sqrt[3]{a} - \sqrt[3]{27a}$ giving your answer in the simplest form. **(4 marks)**

16. The number of people (N) who suffer from Malaria in a month is inversely proportional to the amount of insecticides (M) applied that month. When 5 litres of insecticide are applied, only 1 person suffers from Malaria. Find the equation connecting N and M.

**(3 marks)**

17. P is a set of points (x, y) which satisfies the three inequalities:

$x \geq 0$;

$x + y \leq 4$;

$y \geq x+1$

Show on a graph the region represented by P.

18. Using a ruler and a pair of compasses only, construct a circle of radius 3 cm and a chord **AB** which subtends an angle on the circumference of 45°. Measure the length of the chord.

**(4 marks)**

**Continued**

19. **Figure 6** shows travel graphs of a minibus that leaves Mzuzu at 07:00 hours and arrives in Blantyre at 15:00 hours and a car that leaves Blantyre at 07:30 hours arrives in Mzuzu at 14:30 hours. From Mzuzu, the minibus travels at a constant speed and arrives in Lilongwe at 10:30 hours and immediately proceeds to Blantyre at another constant speed. From Blantyre the car travels at a constant speed and arrives in Lilongwe at 10:30 hours. At noon the car leaves for Mzuzu at a constant speed.

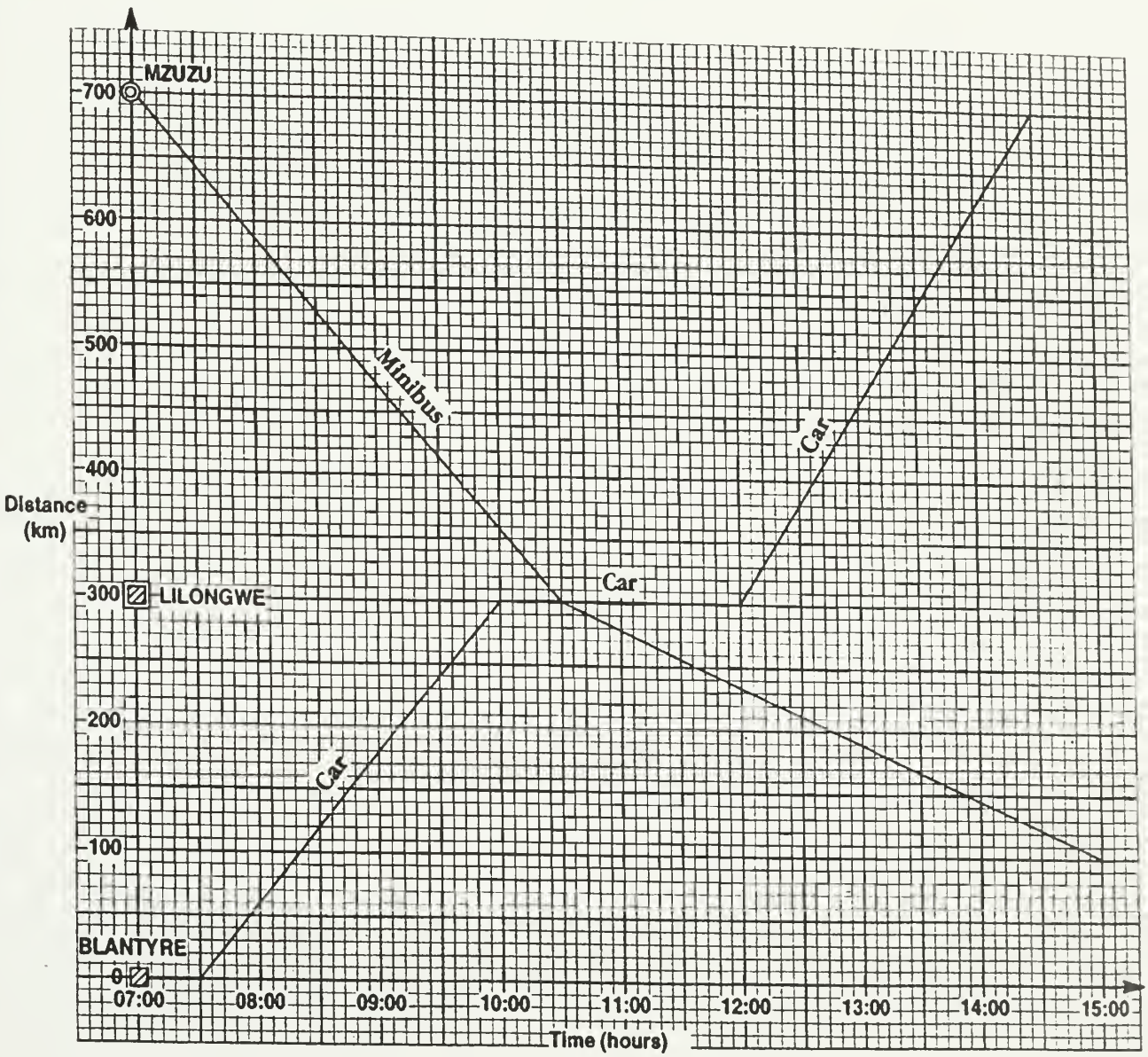



**Figure 6**

Calculate the average speed of the minibus during the time when the car stopped in Lilingwe. **(5 marks)**

Continued ⟶

20. A straight line passes through points A(1, - 1) and B(7, - 9). Calculate the distance between A and B.

**(4 marks)**

21. The volume of a cone is 462 $cm^3$. If its height is 9 cm, calculate its radius.

(Taking $\pi = \frac{22}{7}$, and volume of a cone = $\frac{1}{3}\pi r^2 h$)

**(3 marks)**

22. In **Figure 7**, triangle ABC is similar to triangle BAD.

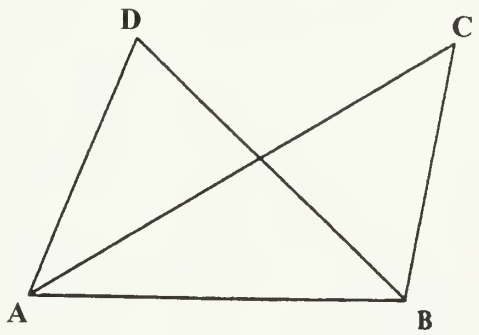



**Figure 7**

If the area of triangle **ABC** = 72 $cm^2$, area of triangle **BAD** = 200 $cm^2$ and **BC** = 6 cm, calculate the length of **AD**.

**(4 marks)**

23. In Figure 8, triangle ABC is isosceles in which AB = AC and angle BAC = 140°.

AB and AC are produced to D and E respectively. The bisectors of angle CBD and angle BCE meet at O.

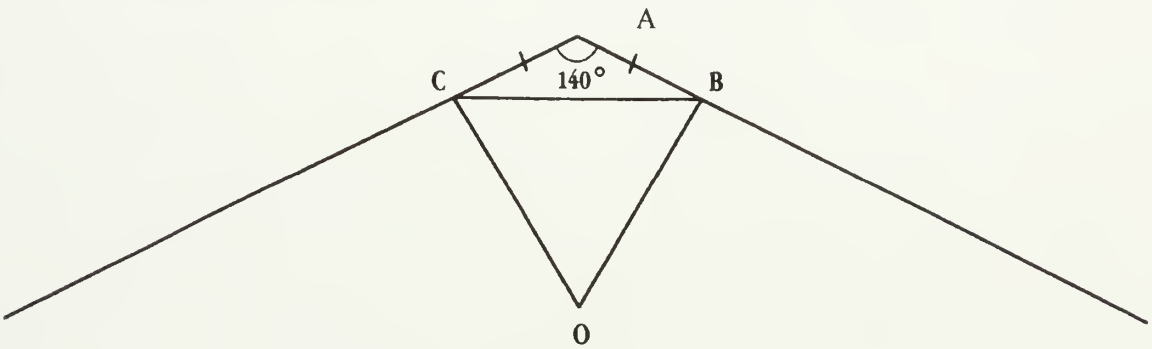



**Figure 7**

Calculate angle **BOC**.

**(6 marks)**

24. When a polynomial $x^3 + kx^2 + x - k$ is divided by $(x - k)$ the remainder is 2. Calculate the value of $k$.

**(4 marks)**

**END OF OUESTION PAPER**

**NB: This paper contains seven**

APPENDIX B

2002 MSCE MATHEMATICS PAPER 2 (REPRODUCED WITH PERMISSION)

**THE MALAWI NATIONAL EXAMINATIONS BOARD**

2002 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

**MATHEMATICS**

Subject Number: M131/II

Wednesday, 16 October

Time Allowed: 2 h 30 mins
1.30 – 4:00 pm

# PAPER II

(100 marks)

**Instructions:**

1. **This paper contains 7 pages. Please check.**
2. Answer **all** the **six** questions in Section **A** and any three questions from Section **B**.
3. The maximum number of marks for each answer is indicated against each question.
4. Mathematical tables, graph paper and answer books are provided.
5. Used graph paper and/or supplementary sheets must be tied together inside the answer book with a string.
6. All working must be clearly shown; It should be done on the same sheet as the rest of the answer.
7. Use of electronic calculators is not allowed in this paper.
8. Write your **Examination Number** on top of each page of your Answer Book.

**© 2002 MANEB**

**Turn over**

**Section A** (55 marks)

Answer **all** the **six** questions in this section.

1. a. Simplify $\dfrac{3\frac{1}{7}[(2\frac{1}{4})^2 - 3\frac{3}{4}]}{3\frac{7}{16}}$ **(4 marks)**

   b. Solve for x, $\log_3 x - 2\log_3 x = 2$ **(4 marks)**

2. a. Simplify $\dfrac{(\sqrt{24}-\sqrt{6})^2}{\sqrt{6}}$ leaving your answer with a rational denominator.

   **(3marks)**

   b. Make $d$ the subject of the formula, $s = \dfrac{n}{2}[2a + (n-1)d]$. **(4 marks)**

3. a. Express as a single fraction

   $$\frac{4}{x-5} - \frac{5}{x(x-5)} - \frac{3}{x}$$

   b. Evaluate $\sqrt[3]{\dfrac{\tan 38^\circ 34'}{35.71}}$ using logarithm tables, leaving your answer to 3 decimal places. **(3 mark)**

4. a. Using a ruler and a pair of compasses only, construct in the same diagram:

   (i) a triangle **ABC** with base **BC** = 12 cm, **AB** = 10 cm and angle **ABC** = $45^o$;

   (ii) a perpendicular from **A** to **BC**, meeting **BC** at **N**;

   (iii) a circle which touches **BC** at **N** and also touches **AC**. Measure the radius of this circle. **(7 marks)**

**Continued** ⟶

b. In **Figure 1**, **XYZ** is a tangent at **Y** on a circle **YSTR.**

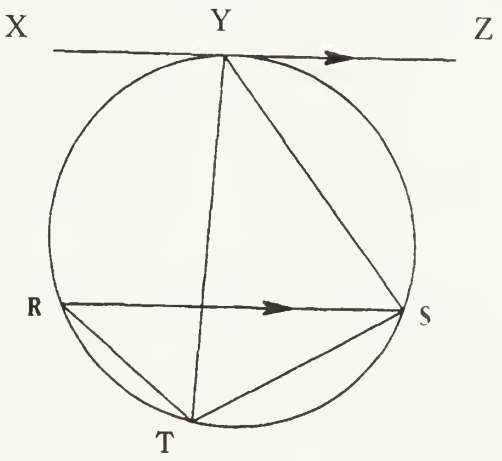



**Figure 1**

If **XZ** is parallel to **RS**, prove that **YT** bisects angle **RTS**.

**(4 marks)**

5. a. A metal bar of length 231 mm and diameter 56 mm is melted down and cast into washers. Each washer is 2 mm thick with an internal diameter of 14 mm and external diameter of 28 mm. Calculate the number of washers obtained assuming no loss of metal.

**(5 marks)**

b. In **Figure 2**, **DCT** is a tangent to the circle **ABC** at **C**.

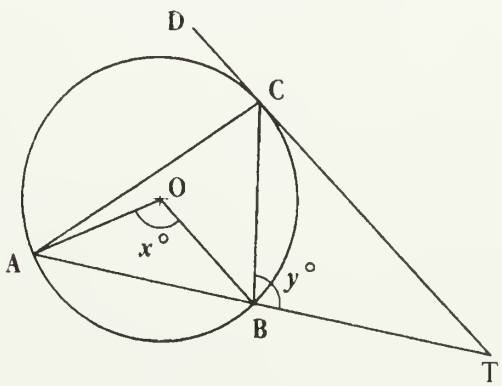



**Figure 2**

Given that angle **CBT** =y °, angle **AOB** = $x°$ and **O** is the centre of the circle, express angle **BCT** in terms of x and y.

**(4 marks)**

Continued →

6. a. In **Figure 3**, **M** is the region bounded by four straight lines.

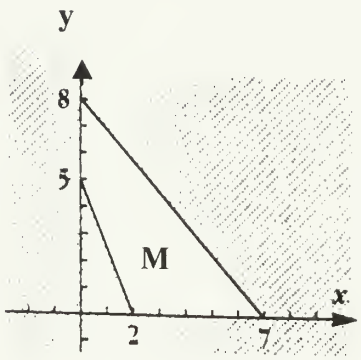



**Figure 3**

Write down the four inequalities describing the region **M**. **(5 marks)**

b. In **Figure 4**, **O** is the centre of the circle **ABC**. The straight line **MOS** is perpendicular to **CA**.

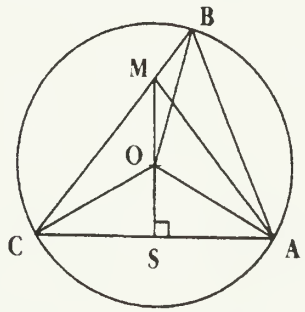



**Figure 4**

Prove that:

(i) the triangles **MCS** and **MAS** are congruent;
(ii) angle **MCO** = angle **MAO**
(iii) **MBAO** is a cyclic quadrilateral

**Section B** (45 marks)

Answer any **three** questions from this section.

7. The worker finds that the earnings per week depend on the time spent in the shop and in the office. If n hours per day are spent in the shop, the weekly earning p kwacha, are given by the relation, $p = 11 + 24n - 3n^2$.

a.(i) Copy and complete the following table of values for $p=11+24n-3n^2$

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| p | 11 | 32 | 47 | 56 | | | 47 | 32 | 11 |

(ii) Using a scale of 2 cm to represent 10 units the vertical axis and 2 cm to represent 1 unit on the horizontal axis, draw the graph of $p=11+24n-3n^2$.

(iii) Use your graph to find the possible times that the worker may stay in the shop to earn K40.00.

**(9 marks)**

b. In the acute-angled triangle ABC, AB = 6 cm, AC = 4 cm and N is the foot of the perpendicular from A to the side BC. Show that $BN^2 - NC^2 = 20$.

**(6 marks)**

8. a. Town X is 10 km due north of town Y. The bearing of ship H from town X is 145° 34'(S 34° 26'E) and town Y is 055° 34'(N 55° 34`E). How far is the ship from Y?

**(6 marks)**

b. The time (T) it takes to enter a stadium is partly constant and partly varies as the number of people (N) entering the stadium. If there are 4 people it takes 12 seconds to enter the stadium and if there are 5 people it takes 14 seconds. How long will it take, if there are 28 people entering the stadium?

**(9 marks)**

9. a. Using a ruler and a pair of compasses only, construct in the same diagram:

(i) a triangle LMN in which angle MLN = 60°, LM = 7.5 cm and LN = 5.0 cm;
(ii) the point R on MN such that MR:MN = 2:1;
(iii) Measure and state the length of LR.

**(8 marks)**

**Continued**

b. **Figure 6** shows the graph of $y = (x + 2)(x-1)(x-3)$

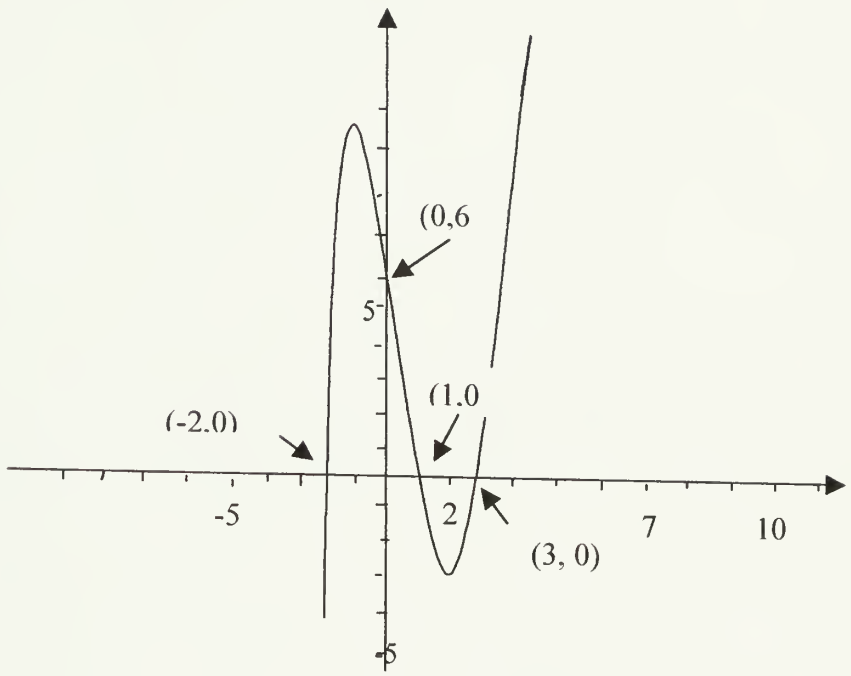



**Figure 6**

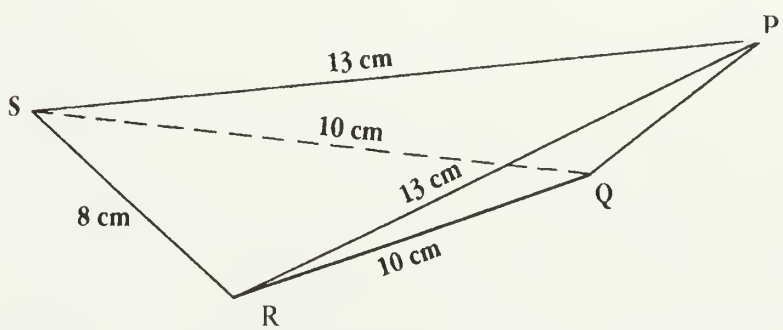
Use the graph to find the solutions of the following equation $x^3 - 2x^2 - 6x - 6 = 0$.

**(7 marks)**

10. a. The sum of the first three positive numbers which are in a GP is 52. The square of the second number is equal to four times the third number. Find the second term of the progression.

**(10 marks)**

b. In Figure 7, PR = PS = 13 cm, QR = QS = 10 cm and RS = 8 cm.




Calculate the angle between the face PRS and base RQS. **(5 marks)**

**Continued→**

11. a. An employee works as a mechanic and a minibus driver at a company. The terms of employment are listed below.

- to work for a maximum of 40 hours per week.
- To spend at least 16 hours per week mending cars and at least 5 hours per week driving minibus.
- to spend at least twice as much time mending cars and driving a minibus.

(i) Express the above conditions as inequalities, using y to represent the number of hours spent mending cars and x to represent number of hours spent driving minibus.

(ii) Using a scale of 2 cm to represent 5 hours on y - axis and 2 cm to represent 5 hours on the x- axis draw graphs of the inequalities and shade the unwanted region.

(iii) If the employee spends 10 hours on driving, use the graph to find the maximum number of hours that can be spent on mending cars.

**(10 marks)**

b. In **Figure 8**, triangle **RST** is such that **RS** = 3 cm, ST = 5 cm and Cosine of angle **RST** =0.600.

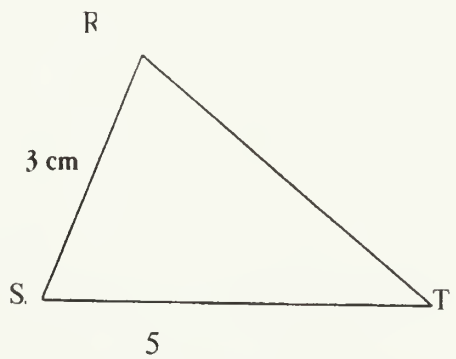



**Figure 8**

Calculate the length of **RT.** **(5 marks)**

12. a. Solve the equation $x^2 + 5x - 1 = 0$. Correct your answer to 2 decimal places. **(7 marks)**

b. WXYZ is a parallelgram. A line through W meets ZY at T and XY produced at U.

(i) Prove that triangles WZT and UWX are similar.

(ii) Given that ZT:TY = 3:2 and the area of triangle WZT = 9 $cm^2$,

Calculate the area of triangle UWX. **(8 marks)**

**END OF QUESTION PAPER**

NB: **This paper contains 7 pages**

2003 MATHEMATICS PAPER 1 (REPRODUCED WITH PERMISSION)

**THE MALAWI NATIONAL EXAMINATIONS BOARD**

2003 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

# MATHEMATICS

**Subject Number: M131/I**

**Friday, 17 October**

**Time Allowed: 2 hours**
**9.00 - 11.00 am**

## PAPER I
(100 marks)

Instructions:

1. **This paper contains 5 pages. Please check.**
2. Answer all the 24 questions in this paper.
3. The maximum number of marks for each answer is indicated against each question.
4. Mathematical tables and answer books are provided.
5. Calculators may be used.
6. Used supplementary sheets must be handed in together with the answer book.
7. All working must be clearly shown; it should be done on the same sheet as the rest of the answer.
8. Write your **Examination Number** on top of each page of your Answer Book.

**© 2003 MANEB**

**Continued/...**

Answer **all** the **twenty four** questions.

1. Factorise completely $x^2 + 3x + 4(x + 3)$. **(3 marks)**

2. Given that $f(x)=x^3-x$, calculate $f(-2)$. **(3 marks)**

3. Express $\frac{3}{\sqrt{2}}$ as a fraction with a rational denominator. **(3 marks)**

4. Given that $a = \begin{pmatrix} 3 & 0 \\ -4 & 4 \end{pmatrix}$ and $b = \begin{pmatrix} 2 & -1 \\ -1 & 0 \end{pmatrix}$, calculate $ab$. **(3 marks)**

5. The universal set $(\varepsilon) = \{10, 20, 30. 40, 50, 60, 70\}$, A = { 10, 30, 60 } and B = {20, 40, 50}, evaluate $A' \text{ I } B$ **(3 marks)**

6. A point **T** has the coordinates $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$. The matrix which transforms T into **T'** is $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$. Calculate the coordinates of **(3 marks)**

7. Calculate vector $\overrightarrow{AB}$ if vectors $A = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ and $B = \begin{pmatrix} -4 \\ 4 \end{pmatrix}$. **(3 marks)**

8. Given $\log_a 2 = 0.6110$ and $\log_a 3 = 0.7039$, calculate $\log_a 6$. **(4 marks)**

9. Calculate the coordinates of the turning point on the curve $y = x^2 + 4x$. **(4 marks)**

10. Express $\frac{1}{x^2 - x - 2} - \frac{1}{x+1}$ as a single fraction. **(4 marks)**

11. Make $m$ the subject of the formula $y = \frac{m}{1+m}$ **(4 marks)**

12. The line joining the points A(3,q), B(5 -q,8) has a gradient of $\frac{1}{2}$. Calculate the value of q. **(4 marks)**

Continued/...

13. Given that $x$ varies jointly as $y$ and inversely as the square of $z$, calculate the missing value in **Table 1**.

**Table 1**

| $x$ | $y$ | $z$ |
|---|---|---|
| 3 | 1 | 2 |
| 1 | 3 | |

**(6 marks)**

14. A box contains 5 red balls, 8 white balls and 7 black balls. If one ball is selected at random, calculate the probability that it is white or black. **(5 marks)**

15. **Table 2** shows the distribution of the number of employees in 43 factories in a town.

**Table 2**

| Number of Employees | 0-39 | 40-59 | 60-79 | 80-99 |
|---|---|---|---|---|
| Number of Factories | 5 | 15 | 13 | 10 |

Draw a histogram.

**(4 marks)**

16. A circle □enter **O** has a tangent **PA** at a point **A. AT** is a chord such that angle **TAP** is acute. If angle **TAP** = 70°, calculate the value of angle **OTA.** **(5 marks)**

17. The first three terms of a G.P. are $x+1$, $x^2-1$, and $(x^2-1)(2x-4)$. Calculate the value of $x$. **(5 marks)**

18. **P** is a set of points $(x,y)$ which satisfies the three inequalities:

$x > -1;$
$y > -2;$
$x + y < 2.$

Using a scale of 2 cm to represent 1 unit on the .x-axis and y-axis draw the region **P.**

**(5 marks)**

Conyinued/...

19. **Figure 1** shows a speed-time graph for a particle during the first 20 seconds of its motion.

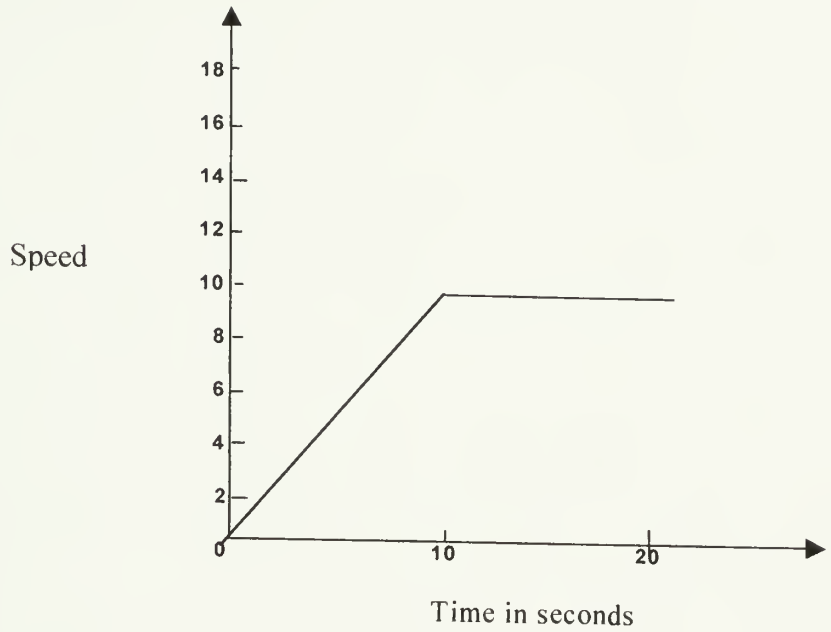



**Figure 1**

Calculate the particle's average speed during the 20 seconds. **(5 marks)**

20. A chord of a circle of radius 5 cm is 8 cm long. Sketch the diagram and calculate the angle subtended by the chord at the □enter of the circle. **(6 marks)**

21. **Figure 2** shows a rectangular box with an open top. The box measures 6 cm long, 2x cm wide and x cm high.

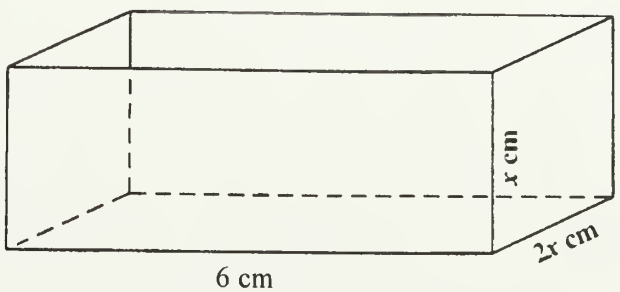



**Figure 2**

Given that the total outer surface area of the box is 108 cm$^2$. Form an equation in $x$ and show that it simplifies to $x^2+6x-27=0$. **(5 marks)**

Continued/...

22. In **Figure 3**, **ABCD** is a trapezium in which **AB** is parallel to **DC**. The diagonals **AC** and **BD** intersect at **O** such that **OD** = **OC**.

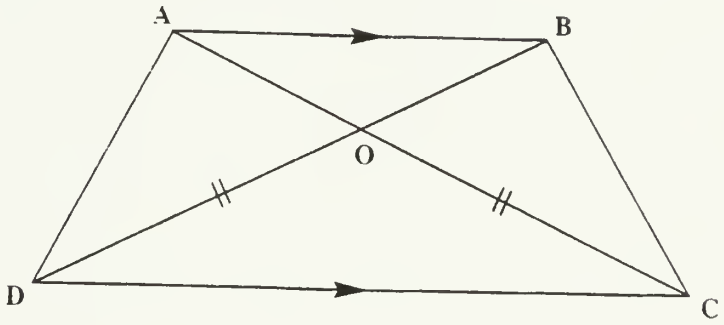



**Figure 3**

Prove that the points A, B, C and D are concyclic. **(4 marks)**

23. Find the remainder when $2x^3-13x^2-8x+12$ is divided by $2x-1$. **(4 marks)**

24. Draw a circle □enter **O** with radius 3 cm. Construct another circle radius 4 cm passing through point **O**. Label its □enter **C**. Label **one** of the intersection points of the two circles **A**. Using a ruler only, construct a tangent to the circle □enter **O** at point **A**.

Measure and state angle **AOC**.

**(5 marks)**

**END OF QUESTION PAPER**

**NB: This paper contains 5 pages**

APPENDIX D

2003 MSCE MATHEMATICS PAPER 2 (REPRODUCED WITH PERMISSION)

**THE MALAWI NATIONAL EXAMINATIONS BOARD**

2003 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

# MATHEMATICS

**Subject Number: M131/II**

**Time Allowed: 2 h 30 mins**

**Thursday, 23 October**

8.00 – 10.30 am

## PAPER II

(100 marks)

**Instructions:**

1. **This paper contains 7 pages. Please check.**
2. Answer **all** the **six** questions in Section A and any three questions from Section B.
3. The maximum number of marks for each answer is indicated against each question.
4. Mathematical tables, graph paper and answer books are provided.
5. Calculators may be used
6. Used graph papers and/or supplementary sheets must be tied together inside the answer book with a string.
7. All working must be clearly shown; it should be done on the same sheet as the rest of the answers.
8. Write your **Examination Number** on top of each page of your Answer Book.

**© 2003 MANEB**

**Turn over**

**Section A** (55 marks)

Answer **All** the **six** questions in this section

1. a. Calculate the value of x if $10^x = 0.01$. **(3 marks)**

   b. In **Figure 1**, **LMN** is a triangle. **D** is a point on **MN** such that angle **LDN** = 90°, angle **LMN** = 60° and angle **NLD** = 45°.

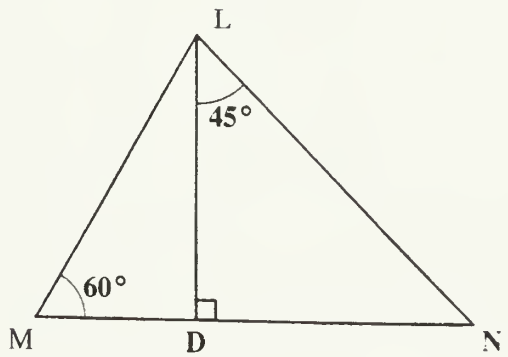


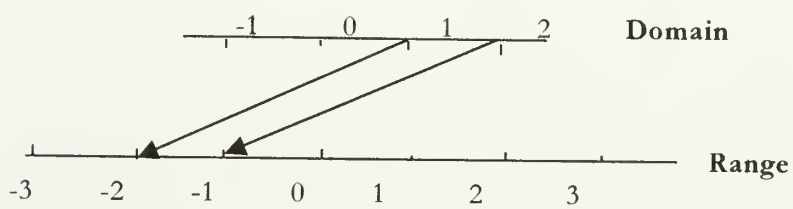

**Figure 1**

Given that **LM** = 12 cm, calculate the length of **DN**. **(6 marks)**

2. a. Given that y is partly constant and partly varies as x, and y = -3 when x = 3, and y = 22 when x = -2, calculate the value of y when x = 2. **(5 marks)**

   b. A and B are two matrices. If $A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}$, find B given that $A^2 = A + B$. **(3 marks)**

3. a. In Figure 2, the function f : x → $x^2$ - 2x - 1 is defined on the domain {-1, *0, 1, 2*}. Copy and complete the mapping diagram for the function.




**Figure**

**(5 marks)**

   b. Make x the subject of the formula

$$y = \frac{p}{2x^3 + q}$$

**(4 marks)**

**Continue→**

4. a. The probability that it rains on a Monday is $\frac{1}{3}$, the probability that the teacher will be present on that day when it rains is $\frac{1}{6}$, and the probability that the teacher will be present when it does not rain is $\frac{7}{10}$.

Draw a tree diagram and label all the probabilities for all the branches **(6 marks)**

b. Factorise completely
$a^2 + 2ab + b^2 - 4$. **(3 marks)**

5. a. In Figure 3, ABCD is a circle and DE is a tangent to the circle at D. AC is parallel to the tangent DE.

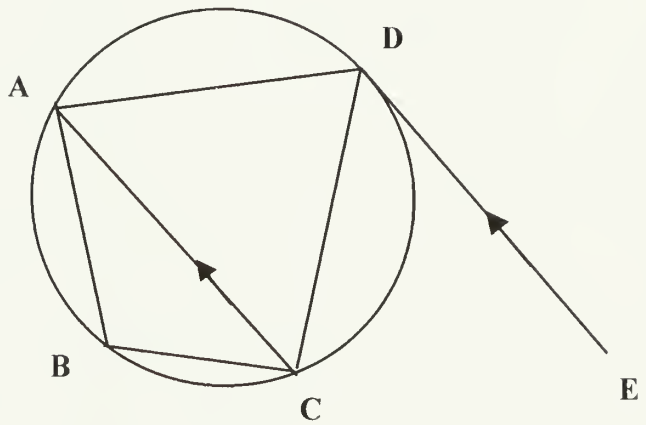



**Figure 3**

Prove that:

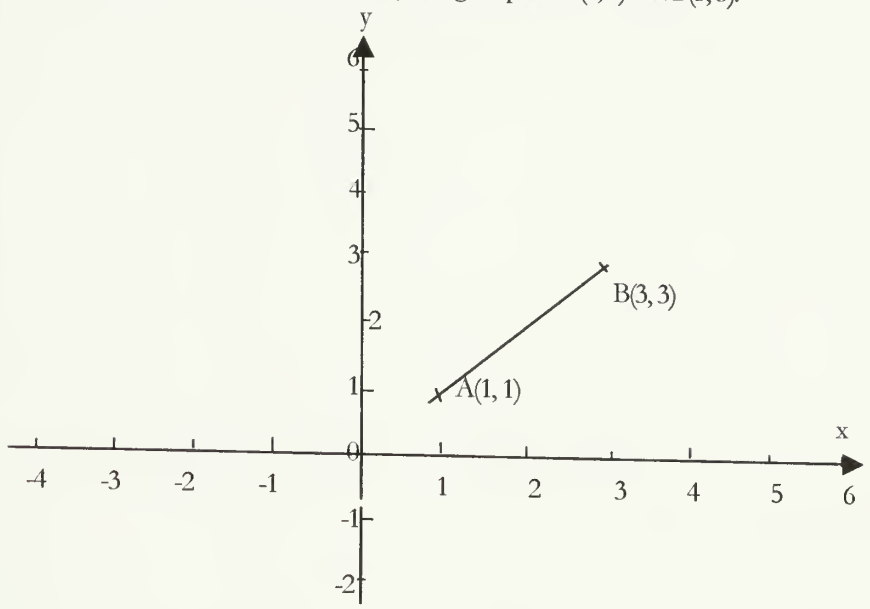
(i) triangle ADC is isosceles:

(ii) angle ABC is twice angle DAC. **(7 marks)**

b. Express $\frac{1}{3\sqrt{2}-3}$ with a rational denominator in its simplest form. **(4 marks)**

Continued

6. a. In Figure 4, AB is the straight line joining the point A(1, 1) and B(3, 3).




(i) Draw the image of **AB** under reflection in the y-axis.

(ii) Find the coordinates of **A'** and **B'**. **(3 marks)**

b. The variance of two temperature measurements, in degrees Celsius, 2 and 2a is 9. Calculate the positive values of a.

**(6 marks)**

**Section B** (45 marks)

Answer any **three** questions from this section.

7. a. Solve for x and y in the vector equation

$$\begin{pmatrix} \frac{x}{3} \\ \frac{y}{3} \end{pmatrix} = 4\begin{pmatrix} -2 \\ 1 \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix}.$$

**(4 marks)**

Continue→

b. A transport company has **three** 30-passenger buses and **nine** 15-passenger buses. The company contracts to transport more than 120 passengers a day to the national parks. It costs K3 000 per day to run each 30-passenger bus and Kl 000 per day to run each 15-passenger bus and the company must spend less than K12 000 per day in order to meet costs.

If x and y are the numbers of 30-passenger and 15-passenger buses used each day,

(i) show that $2x + y > 8$.

(ii) write down **three** other inequalities involving x and y.

(iii) illustrate the solution set of the four inequalities on a graph paper and shade the unwanted regions.

**(11 marks)**

8. a. Solve the simultaneous equations

$$x^2 - y - 5 = 0$$
$$\tfrac{1}{2}y - x = -1$$

**(9 marks)**

b. In **Figure 5**, two quadratic graphs $y = x^2 + 4x - 5$ and $y = \frac{x^2}{2} - 2x + \frac{5}{2}$ are crossing each other at (-5, 0) and (l, 0). **H** is the distance between the maximum of $y = \frac{x^2}{2} - 2x + \frac{5}{2}$ and the minimum of $y = x^2 + 4x - 5$.

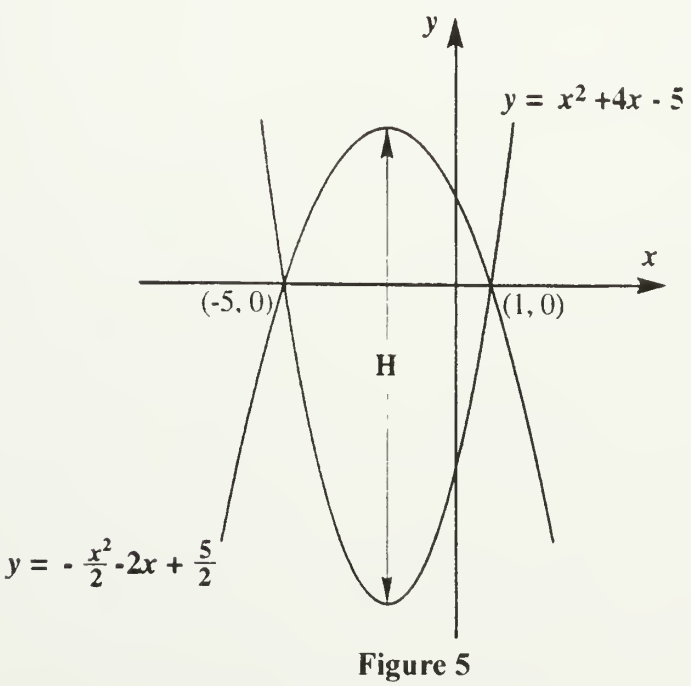



**Figure 5**

**(6 marks)**

Calculate the value of H.

**Continue** ⟶

9. a. A vehicle travels from **P** to **Q** in 8 hours. It starts from rest at **P** increasing its speed steadily to 200 km/h in 2 hours. It then travels at that speed for 1 hour. Finally the vehicle reduces its speed steadily until when it stops at Q, 5 hours later.

(i) Sketch a speed-time graph of the vehicle.
(ii) Using the graph in (i), calculate the distance the vehicle has travelled from P to Q.
**6 marks)**

b. $m^2 - m - 2$ is a factor of $m^3 - 2m^2 - pm + c$. When the polynomial is divided by $m + 2$ the remainder is -12. Find the values of p and c. **(9 marks)**

10. a. Suppose $y = (a + 1)x + 5$ and $y = -2x$ are two parallel straight lines, calculate the value of $a$. **(4 marks)**

b. **Figure 6** represents a solid block made from a right cone and a hemispherical top. The radius of the hemisphere **OB** = 3.5 cm and angle **ACB** = 60°.

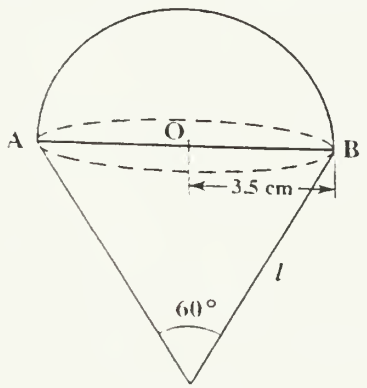
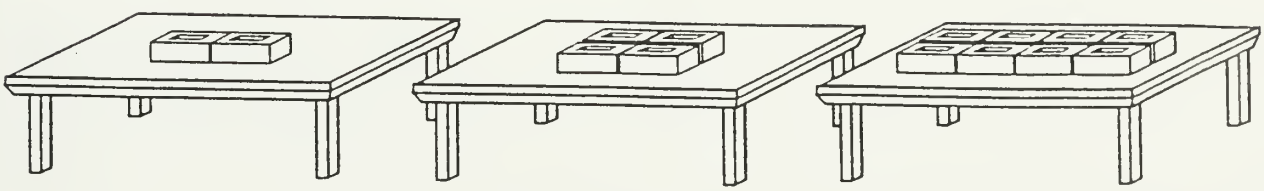



**Figure 6**

Calculate the surface area of the block. (Curved surface area of a cone = $\pi rl$, surface area of a sphere = $4\pi r^2$. Take $\pi = \frac{22}{7}$). **(11 marks)**

11. a. Figure 7, shows the display of new types of bricks laid down on tables. On the 1st table there are 2 bricks, on the 2nd table there are 4 bricks, on the 3rd table there are 8 bricks and so on.



*1st* Table 2nd Table 3rd Table

**Figure 7**

If on the $n^{th}$ table there are 1024 bricks, calculate the value of n. **(6 marks)**

**Continue** →

b. A class of 50 students wrote tests in Mathematics, Biology and Physical Science. The results of the tests were as follows:

- 12 passed Mathematics and Physical Science;
- 19 passed Mathematics and Biology;
- 17 passed Biology and Physical Science;
- 2 passed Physical Science only;
- 5 passed Mathematics only;
- 6 passed Biology only.

If 5 students failed all the three subjects and x passed all the subjects, use a Venn-diagram to calculate the value of x. **(9 marks)**

12. a. The areas of two similar parallelograms are 72 $cm^2$ and 54 $cm^2$. The height of the larger parallelogram is 8 cm. Calculate the corresponding height of the smaller parallelogram.

**(5 marks)**

b. A mini-bus and a lorry left Dedza, at the same time, for Liwonde a distance of 160 km. The mini-bus travelled at an average speed which is 10 km/h faster than the lorry. It arrived at Liwonde 32 minutes earlier than the lorry.

Suppose the average speed of the mini-bus is x km/h,

(i) write down an expression for the time taken by the mini-bus.

(ii) write down an expression for the time taken by the lorry.

(iii) hence, form an equation in x and solve it to find the average speed of the mini-bus.

**(10 marks)**

# END OF QUESTION PAPER

NB: This paper contains 7 pages

APPENDIX E

MSCE SUBJECTS

| OLD CURRICULUM | NEW CURRICULUM |
|---|---|
| Accounts, Principles of | Accounts, Principles of |
| Agriculture | Agriculture |
| Art | Art |
| Bible Knowledge | Bible Knowledge |
| Biology | Biology |
| Chichewa | Business Studies |
| Commercial Studies | Chichewa Language |
| Clothing and Textiles | Chichewa Literature |
| Cookery and Nutrition | Clothing and Textiles |
| English Language | Computer Studies (Unexamined before 2004) |
| French | Cookery and Nutrition |
| General Science | English Language |
| Geography | English Literature |
| Geometrical and Orthographic Drawing | French |
| History | Geography |
| Home Economics | Geometrical and Orthographic Drawing |
| Latin | History |
| Mathematics | Home Economics |
| Mathematics, Addition | Latin |
| Metalwork | Mathematics |
| Physical Science | Mathematics, Addition |
| Woodwork | Metalwork |
| | Physical Science |
| | Science and Technology |
| | Social and Development Studies |
| | Religious and Moral Education (Unexamined before 2004) |
| | Woodwork |

APPENDIX F

2002 ITEM P-VALUES

| Paper 1 | | Paper 2 | |
|---|---|---|---|
| Item | P-value | Item | P-value |
| **1** | **.43** | **1a** | **.45** |
| 2 | .62 | 1b | .09 |
| 3 | .36 | 2a | .19 |
| 4 | .12 | 2b | .23 |
| 5 | .04 | 3a | .34 |
| 6 | .12 | 3b | .24 |
| 7 | .24 | **4a** | **.21** |
| **8** | **.30** | 4b | .12 |
| 9 | .29 | 5a | .04 |
| 10 | .04 | 5b | .06 |
| 11 | .16 | 6a | .08 |
| 12 | .20 | 6b | .22 |
| 13 | .05 | 7a | .52 |
| 14 | .04 | 7b | .15 |
| 15 | .18 | 8a | .10 |
| 16 | .26 | 8b | .36 |
| 17 | .17 | **9a** | **.25** |
| 18 | .16 | 9b | .04 |
| 19 | .09 | 10a | .03 |
| 20 | .07 | 10b | .01 |
| 21 | .19 | 11a | .05 |
| 22 | .09 | 11b | .12 |
| **23** | **.21** | 12a | .27 |
| 24 | .17 | 12b | .14 |

Note: Item p-values for questions from JCE work are those in bold.

APPENDIX G

INVITATION LETTER

The Malawi National Examinations Board
P. O. Box 191
Zomba

30th October, 2003

Mr/Mrs _____
_____
_____
_____

Dear Sir/Madam,

**INVITATION TO A MATHEMATICS STANDARD SETTING WORKSHOP**

I am writing to invite you to a Mathematics Standard Setting workshop to be held at **Chilema** in Zomba from **10th to 13th November, 2003.**

The workshop intends to achieve two related objectives:

1. To define knowledge and skills MSCE candidates should demonstrate in order to be classified as *fail, pass, credit or distinction.*
2. To use the knowledge and skills defined in (1) to set cut scores for various grade categories.

I enclose a copy of the performance level policy definitions which you should study before going to the workshop venue.

You are also encouraged to familiarize yourself with the 2002 and 2003 MSCE Mathematics questions.

Participants are requested to travel by public transport (not coachline) and will be reimbursed their travel expenses on production of relevant travel documents. Those traveling by other means will be reimbursed the equivalent of bus fare.

At the end of the workshop, participants will receive K4000 each.

Please let me know whether or not you will participate.

Yours faithfully,

Dafter J. Khembo

APPENDIX H

MSCE PERFORMANCE LEVEL POLICY DEFINITIONS

| PERFORMANCE LEVEL | DEFINITION |
|---|---|
| **FAILING** (Grade 9) | Students at this level demonstrate a minimal understanding of knowledge and skills in the subject, and have difficulties solving even simple problems. |
| **PASS** (Grades 7 & 8) | Students at this level demonstrate understanding of basic knowledge and skills in the subject, and apply them in limited ways to solve problems. |
| **CREDIT** (Grades 3-6) | Students performing at this level demonstrate understanding of higher-level concepts and skills in the subject, and apply them to solve a wide variety of problems. |
| **DISTINCTION** (Grades 1 & 2) | Students at this level demonstrate broad in-depth understanding of concepts and skills in the subject, and creatively apply them to solve challenging problems. |

APPENDIX I

TIMETABLE FOR THE TRAINING WORKSHOP

| DATE | TIME | ACTIVITY |
|---|---|---|
| Day 1: Monday, 10/10/03 | 8.00 – 8.30 | Introduction and announcements |
| | 8.30 – 10.00 | Lecture on Standard setting |
| | 10.00 – 10.30 | **TEA BREAK** |
| | 10.30 – 12.00 | Presentation of MSCE performance categories and discussion of definitions of MSCE performance categories |
| | 12.00 – 1.30 | **LUNCH BREAK** |
| | 1.30 – 3.00 | Participants develop performance level descriptors and specify knowledge and skill for borderline |
| | 3.00 – 3.30 | **TEA BREAK** |
| | 3.30 – 5.00 | Participants develop performance level descriptors and specify knowledge and skill for borderline |
| Day 2: Tuesday, 10/11/03 | 8.00 –9.00 | Participants set cut scores on sample items, using the descriptors. |
| | 9.00 – 10.00 | Plenary |
| | 10.00 – 10.30 | **TEA BREAK** |
| | 10.30 – 12.00 | Participants take the tests |
| | 12.00 – 1.30 | **LUNCH BREAK** |
| | 1.30 – 3.00 | Participants set cut scores on assigned test papers |
| | 3.00 – 3.30 | **TEA BREAK** |
| | 3.30 – 5.00 | Participants set cut scores on assigned test papers |
| Day 3: Wednesday, 10/12/03 | 8.00 – 10.00 | Participants discuss their results in their sub-panels and revise. |
| | 10.00 – 10.30 | **TEA BREAK** |
| | 10.30 – 12.00 | Participants discuss their results in their sub-panels and revise. |
| | 12.00 – 1.30 | **LUNCH BREAK** |
| | 1.30 – 3.00 | Discussion of impact of cut scores |
| | 3.00 – 3.30 | **TEA BREAK** |
| | 3.30 – 5.00 | Participants reset cut scores |
| Day 4: Thursday, 10/13/03 | 8.00 – 10.00 | Participants reset cut scores |
| | 10.00 – 10.30 | **TEA BREAK** |
| | 10.30 – 12.00 | Participants set final standards |
| | 12.00 – 1.30 | **LUNCH BREAK** |
| | 1.30 – 3.00 | Evaluation and closing |

APPENDIX J

REGISTRATION FORM

Name _____

Sex _____

Age _____

Institution from _____

Teaching experience at MSCE level _____

In which of the following ways are you involved in MSCE work? (you may tick more than one)

a. Teacher

b. Setter

c. Moderator

d. Marker

e. Chief Examiner

f. Curriculum Developer

g. Syllabus Committee member

h. Subject Officer

i. User of MSCE products: College Lecturer, Employer

j. Ministry of Education Official

k. Representative of MANEB management

Date:_____ Signature:_____

APPENDIX K

MSCE MATHEMATICS PERFORMANCE LEVEL DESCRIPTORS

**NUMERATION**

| PASS | CREDIT | DISTINCTION |
|---|---|---|
| Perform basic operations with irrational numbers;<br>Write conjugate surds | Rationalize surd denominators | Multiply 2x2 matrices |
| Perform basic operations with matrices up to multiplication by scalar | Use parallelogram law to add vectors;<br>Use vector method to show collinearity of points | Draw and use speed-time graph to find acceleration and distance covered |
| Understand zero/null and position vectors; find mid-point of a vector | Derive trig ratios of 30,45, 60, 90 degrees;<br>Apply area rule to calculate area and angles of triangles | Perform vector resolution;<br>Find a position vector<br>Demonstrate understanding of parallel vectors; |
| Recall trigonometric ratios;<br>Demonstrate understanding of angles of elevation and depression | Illustrate union and intersection of sets in Venn diagrams | Calculate trig ratios within 0-360 degrees;<br>Solve right triangles using trig ratios;<br>Calculate angles of elevation and depression;<br>Sketch bearing of a point |
| Identify elements in union or intersection of up to three sets | | Use Venn diagrams to analyze and interpret data; |
| | | Some descriptors beyond minimum distinction performance include:<br>Solve problems by applying parallelogram law;<br>Calculate angles in 3-D figures;<br>Solve problems (e.g. calculating bearing of a point) using sine/cosine rules;<br>Solve problems using Venn diagrams. |

## ALGEBRA, PATTERNS, FUNCTIONS

| PASS | CREDIT | DISTINCTION |
|---|---|---|
| Complete square of quadratic expressions | Factorize and calculate roots of quadratic equations;<br>Formulate quadratic equations given roots | Change subject of a formula involving logarithms |
| Change subject of linear equations | Solve simultaneous linear or quadratic equations<br>Change subject of formula involving powers or roots<br><br>Sketch graph of two-variable linear inequality | Apply remainder theorem;<br>Factorize polynomial of third degree, and find their roots;<br>Find coefficients in identical polynomials |
| Construct table of values of a quadratic and cubic functions | Factor and perform basic operations on simple polynomial<br><br>Perform basic operations with algebraic fractions<br><br>Draw and describe results of transformation it in terms of its coordinates;<br>Demonstrate understanding of enlargement | Solve fractional equations<br><br>Draw an enlargement<br><br>Solve logarithmic equations<br><br>Sketch graph of partial variation |
| Find gradient of a line passing through 2 points;<br>Demonstrate understanding of relationship between gradient and parallel lines. | Solve exponential equations;<br><br>Evaluate logarithms of numbers to a given base<br><br>Calculate length of a straight line segment;<br>Determine equation of a straight line given gradient and a point | Determine equation of a line from a graph or passing through two points.<br><br>Find solution of linear programming problem using graph and objective function |
| Write functions in different forms;<br><br>Calculate range given domain and vice versa;<br><br>Draw arrow diagrams | Solve partial variation problems, and calculate their constants<br><br>Graph quadratic and cubic functions;<br>Determine minimum and maximum points of a graph;<br>Find equation of line of symmetry. | Solve quadratic and cubic equations graphically;<br>Formulate quadratic equation given quadratic graph which cuts x-axis;<br><br>Draw an enlargement. |

| | | |
|---|---|---|
| | Calculate nth term, common difference/ratio, and number of terms of AP and GP; Determine sum of AP. | Find sum of GP by formula |
| Demonstrate understanding of a translation | Demonstrate understanding of an enlargement; | Some descriptors beyond minimum distinction performance include:<br>Formulate and solve quadratic and fractional equations from word problems;<br>Apply rules of logarithms in computation;<br>Find equation of line through a given point and parallel to a given line;<br>Solve problems involving joint variation;<br>Illustrate graphically the solution of simultaneous linear inequalities in two variables; Find inequalities in two variables that describe a given region;<br>Illustrate graphically the region described by inequalities; Formulate objective function;<br>Formulate inequalities;<br>Solve simultaneous linear and quadratic or cubic equations graphically;<br>Solve real life problems involving GPs;<br>Find the center of an enlargement |

## STATISTICS AND PROBABILITY

| PASS | CREDIT | DISTINCTION |
|---|---|---|
| Organize data in class intervals and frequencies. Compute mean. | Organize and display data in histogram and frequency polygon | Determine experimental probability of events |
| Determine probability space | Construct probability space table; Solve probability problems using probability space | |
| | | Some descriptors beyond minimum distinction performance include: Calculate variance and standard deviation of ungrouped data; Construct a probability tree diagram Calculate probability of an event using tree diagram. |

## GEOMETRY

| PASS | CREDIT | DISTINCTION |
|---|---|---|
| Demonstrate understanding of chord and angle properties of a circle; Determine angles subtended at center and at circumference by same arc/chord; | Sketch 3-D figures; Find surface area/volume of 3-D figures | Apply chord properties to solve problems; Show that quadrilateral is cyclic; Construct tangents from an external point |
| Identify angles in alternate segments; Show understanding that tangent and radius are perpendicular. | Give formal proofs of designated circle and tangent theorems | |
| State principle of area factor | Use property of proportionality to solve problems in similar figures | |

## GEOMETRY

| PASS | CREDIT | DISTINCTION |
|---|---|---|
| | | Some descriptors beyond minimum distinction performance include:<br>Apply designated tangent theorems to solve problems;<br>Apply chord and circle properties to solve problems |

APPENDIX L

ITEM RATING FORM FOR PAPER 1

(FOR TWO ROUNDS OF RATINGS)

Panelist Name:_____ Date:_____

Subject:_____

| | Pass | | Credit | | Distinction | |
|---|---|---|---|---|---|---|
| Test Item | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | | | | | | |
| 17 | | | | | | |
| 18 | | | | | | |
| 19 | | | | | | |
| 20 | | | | | | |
| 21 | | | | | | |
| 22 | | | | | | |
| 23 | | | | | | |
| 24 | | | | | | |
| Total | | | | | | |

APPENDIX M

ITEM RATING FORM FOR PAPER 2

(FOR TWO ROUNDS OF RATINGS)

Panelist Name:_____ Date:_____

Subject:_____

| | Pass | | Credit | | Distinction | |
|---|---|---|---|---|---|---|
| Test Item | 1 | 2 | 1 | 2 | 1 | 2 |
| 1a | | | | | | |
| 1b | | | | | | |
| 2a | | | | | | |
| 2b | | | | | | |
| 3a | | | | | | |
| 3b | | | | | | |
| 4a | | | | | | |
| 4b | | | | | | |
| 5a | | | | | | |
| 5b | | | | | | |
| 6a | | | | | | |
| 6b | | | | | | |
| 7a | | | | | | |
| 7b | | | | | | |
| 8a | | | | | | |
| 8b | | | | | | |
| 9a | | | | | | |
| 9b | | | | | | |
| 10a | | | | | | |
| 10b | | | | | | |
| 11a | | | | | | |
| 11b | | | | | | |
| 12a | | | | | | |
| 12b | | | | | | |
| Total | | | | | | |

APPENDIX N

## EVALUATION FORM

An edited version of a sample panelist evaluation form from the Handbook for Setting Standards on Performance Assessments by Hambleton, Jaeger, Plake, and Mills (2000a).

MSCE Mathematics Assessment
Standard-Setting Study

Evaluation Form

The purpose of this Evaluation Form is to obtain your opinions about the standard-setting study. Your opinions will provide a basis for evaluating the training and the standard-setting methods.

Please do not put your name on this Evaluation Form. We want your opinions to remain anonymous.
Thank you for taking time to complete this Evaluation Form.

1. We would like your opinions concerning your level of satisfaction with the various components of the standard-setting study. Place a tick (√) in the column that reflects your opinion about the level of satisfaction with the various components of the standard-setting study:

| Component | Not Satisfied | Partially Satisfied | Satisfied | Very Satisfied |
|---|---|---|---|---|
| a. Description of the purposes of MSCE Exams | _____ | _____ | _____ | _____ |
| b. Description of the development of MSCE exams and processing of results | _____ | _____ | _____ | _____ |
| c. Review of the Four Performance Categories | _____ | _____ | _____ | _____ |
| d. Initial Training Activities | _____ | _____ | _____ | _____ |
| e. Practice Exercise | _____ | _____ | _____ | _____ |
| f. Group Discussions | _____ | _____ | _____ | _____ |

In applying the Standard-Setting Method, it was necessary to use definitions of four levels of student performance: Fail, Pass, Credit, Distinction.

2. Please rate the definitions provided during the training for these performance levels in terms of adequacy for standard setting. Please CIRCLE one rating for each performance level.

| Performance Level | Adequacy of the Definition | | | | |
|---|---|---|---|---|---|
| | Totally Inadequate | | | | Totally Adequate |
| Fail | 1 | 2 | 3 | 4 | 5 |
| Pass | 1 | 2 | 3 | 4 | 5 |
| Credit | 1 | 2 | 3 | 4 | 5 |
| Distinctiom | 1 | 2 | 3 | 4 | 5 |

3. How adequate was the training provided on the mathematics test booklet and scoring to prepare you to decide where to place the cut scores? (Circle one)

a. Totally Adequate
b. Adequate
c. Somewhat Adequate
d. Totally Inadequate

4. How would you judge the amount of time spent on training on the mathematics test booklet and scoring in preparing you to decide where to place the cut scores? (Circle one)

a. About right
b. Too little time
c. Too much time

5. Indicate the importance of the following factors in your classifications of student performance.

| Factor | | Not Important | Somewhat Important | Important | Very Important |
|---|---|---|---|---|---|
| a. | The descriptions of Fail, Pass, Credit, Distinction | _____ | _____ | _____ | _____ |
| b. | Your perceptions of the difficulty of the Mathematics Assessment material | _____ | _____ | _____ | _____ |
| c. | Your perceptions of the quality of the student responses | _____ | _____ | _____ | _____ |
| d. | Your own classroom experience | _____ | _____ | _____ | _____ |

e. Your initial rating of the items _____ _____ _____ _____

f. Panel discussions _____ _____ _____ _____

g. The initial ratings of items by other panelists _____ _____ _____ _____

6. How would you judge the time allotted to do the first ratings of the items? (Circle one)

a. About right
b. Too little time
c. Too much time

7. How would you judge the time allotted to discuss the first set of panelists' ratings? (Circle one)

a. About right
b. Too little time
c. Too much time

8. What confidence do you have in the classification of students at the DISTINCTION level? (Circle one)

a. Very High
b. High
c. Medium
d. Low

9. What confidence do you have in the classification of students at the CREDIT level? (Circle one)

a. Very High
b. High
c. Medium
d. Low

10. What confidence do you have in the classification of students at the PASS level? (Circle one)

a. Very High
b. High
c. Medium
d. Low

11. What confidence do you have in the classification of students at the FAIL level? (Circle one)

a. Very High
b. High
c. Medium
d. Low

12. How confident are you that the Standard-Setting Method will produce a suitable set of standards for the performance levels: Pass, Creidt, Distinction? (Circle one)

a. Very Confident
b. Confident
c. Somewhat Confident
d. Not Confident at all

13. How would you judge the suitability of the facilities for our study? (Circle one)

a. Highly Suitable
b. Somewhat Suitable
c. Not Suitable at all

Please answer the following questions about your classification of student performance.

14. What strategy did you use to decide where to place the cut scores?

15. Were there any specific problems that were especially influential in your rating of the items? If so, which ones?

16. Please provide us with your suggestions for ways to improve the standard-setting method.

Thank you very much for completing the Evaluation Form.

# BIBLIOGRAPHY


Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical Report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Educational Statistics.

American Education Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Bennet, J. (1998). *Occasional paper: Setting standards and applying them across different administrations of large-scale high-stakes, curriculum-based public examinations.* Retrieved on January 10, 2003, from http://www.boardofstudies.nsw.edu.au/archives/occasional_papers/occasionalp1_assess.pdf

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56* (1) 137-172

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9,* 215-235.

Beuk, C. H. (1984). A method of reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measuremenet, 21,* 147-152.

Bond, L. (1995). Ensuring fairness in the setting of performance standards. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2*, 311-324. Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and the future. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445-475). Mahwah, NJ: Erlbaum Publishers.

Carson, J. D. (2001). Legal issues in standard setting for licensure and certification. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 427-444). Mahwah, NJ: Erlbaum Publishers.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15(*2), 20-31.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum Publishers.

Collins, B. L. (1995). The consensus process in standards development. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics 2*, 203-219. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). New York: John Wiley & Sons.

Cresswell, M. (2001). *Standard setting methods and issues.* Retrieved on January 11, 2003, from http://www.aea-europe.net/conferences/downloads2001/CresswellKrakow%20Standard%20Setting.ppt

Crocker, L & Algina, J .(1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, & Winston.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliff, NJ: Prentice-Hall.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.), Englewood Cliff, NJ: Prentice-Hall.

Fitzpatrick, A. R., Lee, G., & Gao, F. (2001). Assessing the comparability of school scores across test forms that are not parallel, *Applied Measurement in Education, 14*(3), 285-306.

Giraud, G., Impara, J. C., & Buckendahl, C. (2000). Making the cut in school districts: Alternative methods for setting cut-scores. *Educational Assessment, 6,* 291-304.

Gonzalez, E. J. & Beaton, A. E.(1994). The determination of cut scores for standards. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education* (pp. 171-190). Oxford, England: Pergamon.

Green, B. F. Jr. (2000). *Setting performance standards.* Retrieved January 9, 2003, from www.ipmaac.org/mapac/meetings/2000/berrtgre.pdf

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice 21*(1), 16-22.

Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title 1. In L. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87-114). Washington DC: US Department of Education and the Council of Chief State School Officers.

Hambleton, R. K. (2000). *Translation of NAEP achievement levels to the Voluntary National Tests* (Center for Educational Assessment Research Report No. 397). Amherst, MA: University of Massachusetts.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum Publishers.

Hambleton, R. K. , Brennan, R., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., Van der Linden, W. J., & Zwick, R. (2000). A response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' grading the nation's report card. *Educational Measurement: Issues and Practice,* 5-15.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24(4),* 355-366.

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41-56

Hansche, L. N. (1998). *Handbook for the development of performance standards.* Washington DC: US Department of Education and the Council of Chief State School Officers.

Hau, S. A. (2001). *Major changes in secondary school curriculum.* Paper presented at a symposium on assessment organized by MANEB, Mangochi, Malawi.

Heubert, J. P. & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation,* National Research Council, Washington DC.: National Academy Press.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson and J. S. Helmic (Eds.), *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35* (1), 69-81

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17-31.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed), *Educational measurement* (3rd ed., pp. 485-514). Washington, DC: American Council on Education.

Jaeger, R. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice, 10,* 3-14.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8,* 15-40.

Jaeger, R. M., Cole, J., Irwin, D. M., & Pratto, D. J. (1980). *An interactive structure judgment process for setting passing scores on competency tests applied to the North Carolina high school competency tests in reading and mathematics.* Greensboro, NC: Center for Education Research and Evaluation, University of North Carolina at Greensboro.

Johnson, R. J., Squires, J. R., & Whitney, D. (2002). *Setting the standard for passing professional certification examinations.* Retrieved on January 9, 2003, from www.fma.org/FMAOnline/certifications.pdf

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 3, 425-461.

Kane, M. (1995). Examinee centered vs. task-centered standard setting. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics 2*, 119-140. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Kane, M. T. (1998). Choosing between examinee-centered and test-centered standard – setting methods. *Educational Assessment, 5*(3) 129-145.

Kane, M. T. (2001). So much remains the same: Concepts and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum Publishers.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001) Setting performance standards using body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum Publishers.

Kiplinger, V. L. (1997). *Standard-setting procedures for the specification of performance levels on a standards-based assessment.* Retrieved December 28, 2002, from http://www.cde.state.co.us./cdeassess/asperf.htm

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*(2), 73-96.

Linn, R.L. (1995). The likely impact of performance standards as a function of uses: *From rhetoric to sanctions. Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2*, 367-378. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Livingston, S. A. (1995*).* Standards for reporting the educational achievement of groups. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2*, 39-51. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Livingston, S. A., & Zeiky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, New Jersey: Educational Testing Service.

Loomis, S. C. & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Erlbaum Publishers.

Malawi Certificate Examination and Testing Board (1979). *A short report of the AEB/MCETB comparability study,* Zomba.

Malawi Certificate Examination and Testing Board (1980). *Trends in performance on the individual MCE subjects at credit and distinction levels*, Zomba.

Malawi National Examinations Board (1992). *Mathematics Syllabus: Junior Certificate Mathematics, Malawi School Certificate of Education Mathematics, Malawi School Certificate of Education Additional Mathematics.* Unpublished.

Malawi National Examinations Board (1999). *Malawi School Certificate of Education Examination: Award programme,* Unpublished.

Malunga, L. B. (2000). *Presidential commission of inquiry into the Malawi School Certificate of Education (MSCE) examination results:* A report

Mathews, J. C. (1985). *Examinations: A commentary.* London: Allen & Unwin.

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2,* 221-267. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Mehrens, W. A. & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477-485). Mahwah, NJ: Erlbaum Publishers.

Messick, S (1995). Standards-based score interpretation: Establishing valid grounds for valid inferences. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2,* 291-309. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 219-252). Lincoln, NE: Buros Institute of Mental Measurements.

Mills, C. N. & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. N. Hansche, *Handbook for the development of performance standards (pp. 73-85),* Washington DC: US Department of Education and the Council of Chief State School Officers.

Mills, C. N, Melican, G., & Ahluwilia, N. (1991). Defining minimal competency. *Educational Measurement: Issues and Practice, 10,* 7-10.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum Publishers.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Newton, P. (1997). Examining standards over time. *Research Papers in Education: Policy and Practice, 12* (3), 227-48.

Norcini, J. J. (1990). Equivalent pass/fail decisions. *Journal of Educational Measurement, 27*(1), 59-66.

Norcini, J. J., & Shea, J. A. (1992). Equivalent estimates of borderline group performance in standard setting. *Journal of Educational Measurement, 29*(1), 19-24.

Norcini, J. J. & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10(*1), 39-59.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card.* Washington, DC: National Academy Press.

Phillips, G. W. (1994). Methods and Issues in setting performance standards. In A. C. Tuijnman, & T. N. Postlethwaite (Eds.), *Monitoring the Standards of Education* (pp. 191-212). Oxford, England: Pergamon.

Phillips, S. E. (2001). Legal issues in standard setting for K-12 programs. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 411-426). Mahwah, NJ: Erlbaum Publishers.

Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations.* Unpublished doctoral dissertation, University of Massachusetts.

Pitoniak, M. J., Hambleton, R. K., & Sireci, S. G. (2002). *Advances in standard setting for professional licensure examinations.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum Publishers.

Plake, B.S., Impara, J. C., & Irwin, P. M. (2000). Consistency of Angoff-based predictions of item performance: Evidence of technical quality of results from the Angoff standard setting method. *Journal of Educational Measurement, 37(*4), 347-355.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 201-242). New York, NY: Academic Press.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum Publishers.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspective* (pp. 159-173). Mahwah, NJ: Erlbaum Publishers.

Roeber, E. (2002). *Setting standards on alternative assessments.* National Center of Educational Outcomes. Retrieved on January 13, 2003, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Rudner, L. M. (1992). Reducing errors due to use of judges. *Practical Assessment, Research & Evaluation,* 3(3). Retrieved December 30, 2003 from http://PAREonline.net/getvn.asp?v=3&n=3

Shepard, L. A. (1980). Standard-setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education Evaluation of National Assessment of Educational Progress Achievement Levels. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2,* 143-160. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Shepard, L. A. Glaser, R., Linn, R. L., & Bofrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford, CA: The National Academy of Education.

Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 339-354). Mahwah, NJ: Erlbaum Publishers.

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12,* 301-325.

Stobart, G., Elwood, J., Fordham, R., & Mwanza, J. (1990). *General Certificate of Secondary Education: A comparability study in Geography.* London, England: Joint Council for the GCSE.

Tanner, D. E. (1996). *Decision errors: The trap-door in competency screening.* Retrieved on January 12, 2003, from http://www.coe.ilstu.edu/blnourie/error.htm

Thorn, P., Moody, M., McTighe, J., Kelly, & Peiffer, R. (1990). *Establishing standards for Maryland's School Systems: A systemic approach.* Baltimore, MD: Maryland Department of Education.

Thorndike, R. L. (1982). Item and score conversion by pooled judgments. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 309-317). New York, NY: Academic Press.

Tindall, P. E. N. (1992). *History of Central Africa.* Blantyre: Dzuka Publishing Company.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

Van der Linden, W. J. (1995). A conceptual analysis of standard setting in large-scale assessments. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2,* 97-115. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement, 34(*2), 101-121.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40* (3), 231-253.

*Webster's Universal College Dictionary.* (2001). New York: Random House

Whetton, C., Twist, E., & Sainsbury, M. (2000, April). *National tests and target setting: Maintaining consistent standards.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

Wiley, A. & Guille, R. (2002, April). *The occasion effect for "at-home" Angoff ratings.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Williams, E. B. (1984). *The Malawi examination systems: A report to the World Bank.* Lilongwe

Wolf, A. (1996). Individual choices, incentives and control: understanding assessment dilemmas. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring and selection in international perspectives* (pp. 285-302). Oxford, England: Pergamon.

Zieky, M. J. (1987, November). *Methods of setting standards of performance on criterion referenced tests.* Paper presented at the 13th International Conference of the International Association for Educational Assessment, Bangkok.

Zieky, M. J. (1995). A historic perspective on setting standards. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessments Governing Board and the National Center for Education Statistics, 2*, 1-38. Washington, DC.: National Assessment Governing Board and National Center for Education Statistics.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-51). Mahwah, NJ: Erlbaum Publishers.

Zieky, M. J. & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests.* Princeton, New Jersey: Educational Testing Service.

Zoani, A. J. R. (1989). *Application of standard setting methods in public examinations.* Unpublished doctoral thesis, Monash University.